

Délégation aux Systèmes d'Information
Pôle des "services d'appui à la recherche"
Informatique scientifique

Appel à projets interne

SPIRALES 2008

Formulaire de demande DSI-SPIRALES

« Soutien aux Projets Informatiques dans les Equipes Scientifiques »

Remise des projets :
16 novembre 2007
à spirales@ird.fr

Contact :

Régis Hocdé - Informatique Scientifique

regis.hocde@ird.fr

ou equipe-is@ird.fr

Demande d'un soutien DSI sur les projets informatiques des UR/US.

Formulaire de demande DSI-SPIRALES 2008
« Soutien aux Projets Informatiques dans les Equipes Scientifiques ».

Le présent formulaire comporte différentes parties qui doivent ou non être renseignées selon la nature de votre projet. La modification du formulaire en une série de questions précises est destinée à **faciliter le travail des évaluateurs**.

Les propositions doivent être adressée sous forme électronique (au format RTF, DOC ou PDF) à l'adresse suivante : spirales@ird.fr

1. Nature du projet

Cette partie (questions 1 à 4) doit être renseignée **quelque soit la nature de la proposition** (*nouveau projet ou continuum d'un projet SPIRALES existant, étude de faisabilité, projet finalisé de développement d'une application IS ou autre*).

1. Titre du projet :

Développement de la plateforme bio-informatique IRD plante (EST-DB) : Analyse, conception et développement de nouveaux modules pour l'annotation automatique et pour la génomique comparative.

2. Résumé du projet proposé (5 lignes maximum)

Le développement croissant de projets génomiques plante à l'IRD a conduit les chercheurs des UMRs concernées à mutualiser leurs moyens et à grouper leurs efforts pour mettre en place une plateforme bioinformatique plante à l'IRD (Resp. C. Tranchant-Dubreuil). Depuis 2001, une chaîne ou pipeline d'annotation couplée à une base de données (Application EST-db) ont été conçus pour analyser, exploiter et stocker les données génomiques. Cet outil est en constante évolution et c'est dans ce contexte que s'inscrit ce projet SPIRALE qui propose une optimisation de l'outil existant couplé à l'ajout de nouvelles fonctions.

3. Type de projet

☐ **Nouveau projet SPIRALES :**

☐ **Etude de faisabilité** : **Demande d'appui pour une analyse fine des besoins** et la formulation de spécifications, éventuellement développement d'un prototype (en vue d'une seconde phase destinée au développement et à la réalisation du projet),

Ne renseigner que les parties 1 – 2 – 3 – 5 du présent formulaire.

☐ **Projet finalisé de développement d'une application IS** (proposition finalisée et détaillée en matière d'expression des besoins, d'identification des solutions et des moyens…),

Ne renseigner que les parties 1 – 2 – 3 – 5 – 6 – 8 du présent formulaire.

Joindre le cas échéant tous documents utiles (document de spécifications techniques et fonctionnelles, cahier des charges, propositions techniques et financières reçues…)

☐ Projet **autre qu'un développement** d'application IS (proposition finalisée concernant tous autres domaines : animations, évènements, traitement de données, calcul intensif…),

Ne renseigner que les parties 1 – 2 – 3 – 5 – 7 – 8 du présent formulaire.

☒ **Continuum d'un projet SPIRALES existant** (prévu sur 2007-2008 ou suite d'un précédent projet SPIRALES)

■ Continuum d'un **développement d'une application IS**,

Ne renseigner que les parties 1 – 2 – 3 – 4 – 6 – 8 du présent formulaire.

☐ Continuum d'un projet **autre qu'un développement** d'application IS,

Ne renseigner que les parties 1 – 2 – 3 – 4 –7 – 8 du présent formulaire.

> Les demandes d'hébergement d'applications IS, d'accès à un serveur de développement, de création de dépôt Subversion (SVN), de formations IS… ne constituent pas des demandes SPIRALES et doivent être adressées directement à equipe-is@ird.fr sans échéance particulière.

4. Durée prévue :

☐ Durée prévue : ☐ 1 an ■ 2 ans

Pour les continuums : date de démarrage du projet : o 2004 o 2005 o 2006 ■ 2007

2. Porteur(s) de projet

> Cette partie (questions 5 à 17) doit être renseignée **quelque soit la nature de la proposition** (*nouveau projet ou continuum d'un projet SPIRALES existant, étude de faisabilité, projet finalisé de développement d'une application IS ou autre*).

5. Unité :

■ UMR ☐ UR ☐ US N° 188 Nom : DIA-PC

6. Département

☐ DME ■ DRV ☐ DSS

7. Nom du porteur de projet :

Christine Dubreuil-Tranchant et Valérie Hocher

8. Statut et coordonnées du porteur de projet

Christine DUBREUIL-TRANCHANT	IE	IRD Montpellier	0467416334	0467416222	christine.tranchant@mpl.ird.fr
Valérie HOCHER	CR1	IRD Montpellier	0467416196	0467416222	hocher@mpl.ird.fr

9. Nom et coordonnées du Directeur d'Unité (*si différent*) :

Serge HAMON	DR1	IRD Montpellier	0467416194	0467416222	hamon@mpl.ird.fr

10. Aval du directeur d'unité (*obligatoire*).

Le travail effectué au niveau du projet EST-dbde part son aspect collaboratif et les modules déjà opérationnels constitue déjà en soi un ensemble remarquable. Il a permis à 4 Unités de Recherche de se fédérer autour d'une approche DSI Spirale, pour concevoir et développer un produit tout à fait original. Il reste cependant quelques petits aménagements à terminer en particulier au niveau du fonctionnement automatiser et des visuels. Il s'agit maintenant d'aller jusqu'au bout de l'initiative et de finaliser l'application via une documentation technique et un manuel d'installation – documentation en anglais. Enfin, il me semble évident que l'intégration d'un module connecté – et donc se servant de modèle – à la référence mondiale "Gene Ontology" constitue un plus indéniable qui donnera à ce produit une valeur internationale incontestée. Je donne donc un avis très favorable pour ce projet.

Serge HAMON – DU – UMR188

11. Implantation principale de l'unité :

Centre IRD de Montpellier

12. Site de déroulement du projet :

Centre IRD de Montpellier

13. Site administratif à partir duquel se feront les dépenses budgétaires

Centre IRD de Montpellier

14. Projets inter-unité ou inter-organismes :

☐ **Projet inter-unités** ☐ **Projet inter-organismes**

15. Liste des unités ou organismes partenaires du projet

IRD	UMR DIA-PC	S. Hamon	IRD Montpellier
IRD	UMR 186 RPB	M. Nicole	IRD Montpellier
IRD	UR 192 Palmiers	J Tregear	IRD Montpellier
IRD	UR 060/Clifa	H. Chrestin	Univ Mahidol, Bangkok Thaïlande
IRD	UMR LSTM	B. Dreyfus	IRD Montpellier Baillarguet
INRA	UMR DIA-PC	S Hamon	INRA Mauguio
CNRS/ Univ Lyon1	UMR 5557	R. Bally	Université Lyon 1

16. Liste des intervenants impliqués de manière effective dans la réalisation du projet :

(autant de fois que nécessaire)

Christine DUBREUIL-TRANCHANT	IE	UMR DIA-PC	IRD Montpellier	Chef de projet. Encadrement bio-informatique
Valérie HOCHER	CR1	UMR DIA-PC	IRD Montpellier	Création d'un groupe de travail (Animateur V. Hocher) pour l'encadrement et le suivi du projet pour les aspects biologiques.
Florence AUGUY	IE	UMR DIA-PC	IRD Montpellier	
Perla HAMON	DR2	UMR DIA-PC	IRD Montpellier	
Alexandre DE KOCHKO	DR2	UMR DIA-PC	IRD Montpellier	
Joëlle RONFORT	DR2	UMR DIA-PC	INRA Mauguio	
Diana FERNANDEZ	CR1	UMR 186 RPB	IRD Montpellier	
Fabienne MORCILLO	CR	UR 192 Palmiers	IRD Montpellier	
Tim TRANBARGER	CR2	UR 192 Palmiers	IRD Montpellier	
Hervé CHRESTIN	DR2	UR 060/Clifa	IRD-Thaïlande	
Fabienne CARTIEAUX	CR2	UMR LSTM	IRD Montpellier	
Philippe NORMAND	DR	UMR 5557	Univ. Lyon 1	

Ce projet est le continuum d'un projet démarré en 2007. Le cahier des charges établi en 2007 (dont nous présenterons le bilan) est réalisé via un prestataire de service en bioinformatique. 2 intervenants sont impliqués.

Isabelle VERGELY		Société ASA	Chef de projet	
Benoît PARRA		Société ASA	Développeur	

17. Disponibilité / implication de chacun des intervenants effectifs : exprimée en % de temps-homme ou en jours-homme (ETP total ou pour une période)

Ex : Dupont – forte disponibilité - 50% ETP sur la durée du projet

Martin – très faible disponibilité – 0,5 jour / mois

L'essentiel est de donner un ordre de grandeur (et non pas une évaluation monétaire) : s'agit-il de 4 jours de travail (4 jours ETP) pour l'année, 15 jours ETP ou 40 jours ETP (un jour par semaine) ou de s'impliquer à temps complet (200 jours ETP)… ?

Nom	Organisme	Disponibilité	Implication
Christine DUBREUIL-TRANCHANT(*)	IRD	Forte disponibilité	1 jour / sem
Valérie HOCHER	IRD	Moyenne disponibilité	0,5 jour/ sem
Florence AUGUY	IRD	Faible disponibilité	1 jour/ mois
Perla HAMON	IRD	Faible disponibilité	1 jour/ mois
Alexandre DE KOCHKO	IRD	Faible disponibilité	1 jour/ mois
Joëlle RONFORT	INRA	Faible disponibilité	1 jour/ mois
Diana FERNANDEZ	IRD	Faible disponibilité	1 jour/ mois
Fabienne MORCILLO	CIRAD	Faible disponibilité	1 jour/ mois
Tim TRANBARGER	IRD	Faible disponibilité	1 jour/ mois
Hervé CHRESTIN	IRD	Faible disponibilité	1 jour/ mois
Fabienne CARTIEAUX	IRD	Faible disponibilité	1 jour/ mois
Philippe NORMAND	CNRS	Faible disponibilité	1 jour/ mois

* Depuis Août 2007, C. Dubreuil-Tranchant est en congés maternité et est remplacée par Olga PLECHAKOVA jusqu'à son retour prévu en Janvier.

3. Moyens / appui demandés à la DSI

Cette partie (questions 18 à 27) doit être renseignée **quelque soit la nature de la proposition** (*nouveau projet ou continuum d'un projet SPIRALES existant, étude de faisabilité, projet finalisé de développement d'une application IS ou autre*).

18. Contribution demandée à la DSI pour 2008 en euros HT et TTC :

Montant 2008 demandé : **20000 € HT soit 23920 € TTC (pour les projets en France)**

Ventilation par poste :

Fonctionnement :

Equipement :

Prestation de service : 24000 euros

19. Demande envisagée pour 2009 – si projet de 2 ans - en euros HT et TTC :

20. Montant(s) précédemment attribué(s) par la DSI - en euros HT :

	2004	2005	2006	2007
Montants attribués (€ HT)				**24000 €**

21. Moyens affectés au projet et Cofinancements acquis hors SPIRALES (€ HT) :

Autres sources de financements acquis : Montant (€ HT) :

Moyens apportés par l'unité (hors ressources humaines) Montant (€ HT) : 10 000 € en 2007

Moyens demandés par l'unité pour 2008 : Montant (€ HT) : 10 000 €

22. Moyens humains affectés au projet :

Total des moyens humains affectés au projet par les unités et partenaires (exprimé en total de jours-homme ou ETP (Equivalent Temps Plein)) (*cf. définition et exemple à la question 17.*) :

- Un IE responsable du projet qui assure l'encadrement bio-informatique (C. Dubreuil-Tranchant ou remplaçant).
- Groupe de travail (Animateur V. Hocher) composé de chercheurs pour l'encadrement biologique. (expression des besoins des chercheurs, suivi du projet, test des outils mis en place).

Soit 6 mois équivalent temps plein

- Prestataire de service : 2 personnes pendant 6 mois.
  - 1 chef de projet
  - 1 développeur

23. Coût total estimé du projet (toutes années confondues) :

Estimation du coût total du projet toutes années SPIRALES confondues : crédits SPIRALES, moyens fournis par l'unité et cofinancements acquis (hors ressources humaines) : 60000 € HT

24. Ressources humaines extérieures mobilisées ou demandées:

☐ Compétences mobilisées ou souhaitées (profil type) :

☒ Intervention d'un/de prestataire(s) de service :

o Mobilisation d'un/de stagiaire(s) *(sous réserve de compétences fortes en informatique scientifique au sein de l'équipe porteur du projet et de capacités de l'équipe à dégager du temps pour assurer un réel encadrement)*

☒ Demande d'appui de l'équipe 'Informatique scientifique' de la DSI / pour l'appui méthodologique et le suivi de projet :

☐ Demande d'appui de l'équipe 'Informatique scientifique' de la DSI / pour le développement et/ou la réalisation du projet (avec estimation du temps-homme nécessaire) :

*La DSI, suite au comité d'évaluation, pourra pour quelques projets et sur quelques sites (Nouméa, Dakar, Montpellier…) et dans la limite des **moyens humains de la DSI disponibles**, convertir ces demandes d'appui ou de financement de prestataire de service en temps-homme, c'est-à-dire par une intervention directe du 'pool informatique scientifique'.*

25. Demande d'un dépôt Subversion (SVN) :

Description des besoins pour ce projet SPIRALES (une demande formelle et détaillée, avec signature de la charte sera néanmoins nécessaire dans un 2nd temps) - (Définition SVN: http://fr.wikipedia.org/wiki/Subversion_(logiciel))

26. Demande d'hébergement(s) / d'accès à un (des) serveur(s) 1/ de développement et de tests pour la durée du projet, 2/ de 'pré production' et de recette pendant ou à l'issue du projet, 3/ d'exploitation à l'issue du projet :

Description des besoins pour ce projet SPIRALES : technologies, capacité… (une demande formelle et détaillée, avec

signature de la charte sera néanmoins nécessaire dans un 2nd temps)

L'application sera hébergée sur le serveur de production de la plate-forme bio-informatique de Montpellier.

27. Appui de la DSI apporté pour l'élaboration du projet ?

Si vous avez bénéficié de l'appui de la DSI (coordination IS, pool d'informaticiens scientifiques de Dakar ou Nouméa, SIL…) pour l'élaboration de cette proposition, décrivez très brièvement le type d'appui.

Comme pour tout projet bio-informatique, le SIL de Montpellier est impliqué (expertise, conseil technique et administration système de la plate-forme bio-informatique de Montpellier).

4. Bilan / Etat d'avancement des phases précédentes (seulement pour les demandes de continuums)

Cette partie (questions 28 à 32) **ne concerne que les demandes de continuums** pour des projets SPIRALES initiés au cours des années précédentes.

Il est vivement conseillé d'**accompagner la demande de tous documents utiles** :

rapport de phases préliminaires, cahier des charges, résultats, prototype, 'vues écrans' de l'application développée, démonstrateur en ligne…

28. Etat d'avancement du projet :

A- Acceptation du projet : début 2007

Pour la réalisation de ce projet, il a été décidé de passer par un prestataire de service.

Avril et Mai 2007 :	Rédaction du cahier des charges précis par Christine Dubreuil-Tranchant et validation par le groupe de travail.
Juin 2007 :	Identification de prestataires de service et mise en concurrence Réalisation des études par les différents prestataires Adaptation du cahier des charges
Fin Juin 2007 :	Société ASA (Advanced Solution Accelerator, Castelnau Le Lez) retenue comme prestataire par le groupe de travail. Finalisation du cahier des charges et de l'échéancier
Juillet 2007 :	Début de la prestation

A noter : 1/07/07 – 15/07/07 : 15 jours d'échange entre I Vergely et C Dubreuil-Tranchant pour expliquer le fonctionnement d'EST-DB, car C. Tranchant devait partir en congé maternité et ne pouvait assurer l'encadrement d'I. Vergely (ASA) après le 1/08/07.

En documents joints : le cahier des charge IRD, l'étude d'ASA et l'échéancier initial.

B- Les différentes étapes du projet :

Livrable 1 : Finaliser le développement de l'application EST-db.

- Documentation du code en anglais ;
- Documentation technique, infrastructure de l'application en anglais ;
- Manuel d'installation

Livrable 2 : Améliorer la version actuelle de l'application EST-db.

- Interface de consultation pour les clusters,
- Gestion des identifiants des contigues lors de la mise à jour d'un projet de clusterisation;

Livrable 3 : Développement et intégration d'un nouveau module d'analyse pour « l'annotation fonctionnelle » automatique des séquences avec une base de données contenant la nomenclature « Gene Ontology » (GO).

3a : Conception : définition précise du fonctionnement du module.
- Veille technologique des outils existants dans le domaine public :
  - o Outils d'annotation ;
  - o Bases de données disposant de la nomenclature GO ;
- Comparaison des outils trouvés afin d'identifier le plus adapté pour le pipeline existant ;
- Définition des éléments résultant du processus d'annotation devant être inclus dans la base de données EST-db;

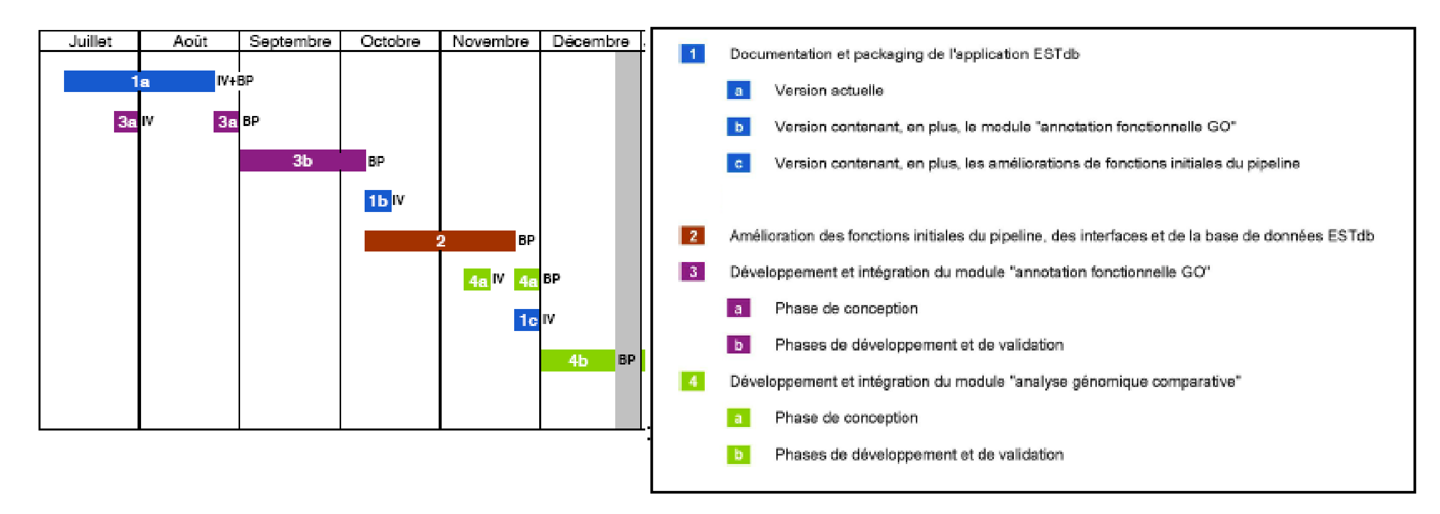
3b : Phase de développement et de validation pour :
- Phases de conception, de développement pour :
  - o Le fonctionnement du module ;
  - o Le stockage de résultats dans la base de données EST-db;
  - o La consultation de ces informations ;
  - o Test du module par les utilisateurs et intégration de ce module au pipeline

Livrable 4 : Développement et intégration d'un nouveau module « analyse génomique comparative ».

4a : Conception : définition précise du fonctionnement du module.
- Veille technologique des outils existants dans le domaine public :
- Mode de visualisation/présentation des résultats ;
- Eléments résultant de l'analyse devant être inclus dans la base de données EST-db;

Partie non encore faite. Devrait être faite pour fin

29. Respect de l'échéancier (et rappel de l'échéancier) :



Un échéancier initial a été suivi. Suite au départ de C. Tranchant (début Août), nous avons rencontré quelques difficultés notamment pour des autorisations d'accès. Par ailleurs, l'étude du module Gene Ontology a demandé plus de temps que nécessaire afin d'ajuster les possibilités techniques aux demandes des utilisateurs.

Ces deux évènements ont retardé d'environ 1 mois l'échéancier initial. Nous avons donc modifié cet échéancier : nous avons allégé certains livrables et reporté une partie du travail pour 2008.

30. Objectifs atteints OU non encore réalisés (et raisons) :

Au 15/11/07, les livrables suivants ont été réalisés :

La documentation technique a été rédigée en anglais (Livrable 1) ; elle sera finalisée en décembre 2007

Le module « Annotation Gene Ontology » a été implémenté (Livrable 3 A et B)

Le module « Génomique Comparative » (Livrable 4) : en cours de conception. La phase de développemnt est reportée sur 2008 suite aux différents retards.

L'amélioration de la version actuelle d'EST-db(Livrable 2) : La gestion des mise à jour de clusterisation est en cours. Les autres points demandés sont reportés sur 2008 suite au retard expliqué précédemment.

31. Livrables produits (outils, documentations, méthodes, URLs…) / fournis à l'équipe IS :

L'objectif du projet est d'obtenir un outil, EST-db, transférable et utilisable. Il sera donc fourni à l'équipe IS dès :

1/ sa finalisation et sa validation par C. Dubreuil-Tranchant et le groupe de travail

2/ son dépôt à l'Agence de Protection des Logiciels

3/ la validation de l'autorisation de diffusion par les DU des UMRs IRD impliquées.

32. Moyens humains et services de la DSI mis à contribution en 2007 (personnes ressources du pool 'IS', dépôt SVN, hébergement sur serveur, formation…) ?

Y. Pournin, SIL

B. Granouillac DSI

5. Description des besoins

Cette partie (questions 33 à 36) concerne tous les **nouveaux projets (étude de faisabilité, projet finalisé de développement d'une application IS ou autre).**

La demande peut-être être **accompagnée de tous documents utiles** :

présentation du projet global ou descriptif du projet, rapport de phases préliminaires, étude de faisabilité, dossier d'expression des besoins ou ***cahier des charges****, devis détaillé…*

33. Objectifs scientifique

Ce projet est un continuum de 2007. Il a donc déjà été décrit, mais pour des questions de commodités (relecture possible par les évaluateurs, …) nous avons re-copié certaines parties du projet initial. Elles sont surlignées en gris.

A- Situation générale du projet (Copie de 2007)

Depuis plusieurs années, les équipes plantes IRD appartenant à différentes UMRs/URs développent des projets de génomique concernant des plantes tropicales d'intérêts majeurs pour les pays du SUD (*Coffea*, *Hevea*, Arécacées, Casuarinacées, ...) et s'articulent essentiellement autour de 2 plantes modèles entièrement séquencées *Arabidopsis thaliana* et *Oryza sativa*. La production en masse de données génomiques et protéomiques et la nécessité de leur analyse a conduit à la mise en place d'une plateforme bioinformatique plante à l'IRD (Resp. C. Tranchant-Dubreuil).

Une chaîne ou pipeline d'annotation d'ESTs utilisant différents logiciels bio-informatiques gratuits (Blast, Stackpack...) conjugués à des algorithmes puissants est maintenant disponible. Les données brutes placées en entrée du pipeline ainsi que les données produites par le pipeline sont stockées dans une base de données (EST-db), ce qui permet de gérer et d'exploiter les données produites sur les différentes plantes tropicales. Différentes équipes ont déjà valoriser l'utilisation du pipeline au travers de publications scientifiques (Voir Paragraphe « Liste des Publications »).

Plusieurs équipes IRD plante se sont maintenant engagées dans des projets de génomique comparative et la réalisation de ces programmes nécessite **le développement d'une nouvelle chaîne de traitement du pipeline** combinant les logiciels bioinformatiques adéquats (correspondant à une nouvelle fonctionnalité de l'application EST-db). Ces projets concernent en priorité les différentes espèces propres à l'IRD et s'étendront à des espèces travaillées dans des équipes d'autres organismes (INRA, CNRS, ...) associées à l'IRD par le biais des UMRs ou de projet de recherche communs. Une bréve description des

programmes de recherche concernés est proposée (Voir paragraphe « Descriptif des Projets Scientifiques») et permettra de mieux mesurer la nécessité de développer cette nouvelle fonctionnalité.

Cependant, le développement de ce nouveau module est conditionné par **l'amélioration préalable du pipeline existant** afin, d'une part, de le rendre convivial et générique et, d'autre part, d'y ajouter certaines fonctionnalités, telle que l'annotation automatique des séquences en grandes fonctions selon Gene Ontology ou encore l'affichage correct des séquence et de leur contig. Enfin, pour **valoriser l'outil EST-db**, une étape de documentation est indispensable pour sa **déclaration à l'Agence de Protection des Logiciels et sa distribution** (plusieurs partenaires souhaiteraient l'installer et l'utiliser sur leur plate-forme).

L'objectif de ce projet SPIRALE est de réaliser ces différentes opérations afin d'obtenir un outil EST-db convivial, générique, facilement transférable à d'autres équipes (IRD ou extérieure) et répondant aux demandes actuelles des projets scientifiques (traitements de données de séquençage, analyses comparatives).

B- Situation du projet pour 2008:

Suite à l'acceptation de ce projet, un prestaire de service a été engagé (ASA) et travaille actuellement sur les différents éléments définis par le cahier des charges. Le projet a relativement bien évolué et nous souhaiton sle continuer en 2008 afin de le finaliser.

Quelques modifications sont à noter : un programme scientifique supplémentaire (Cf Programme 2) a été ajouté suite à l'obtention d'un projet Génoscope. Par ailleurs, l'équipe IRD travaillant sur Nouméa en association avec l'Université souhaite installer EST-DBle plus rapidement possible et constitue avec le programme Hévéa en Thaïlande et le programme Café à la Réunion une demande très forte de transfert de l'outil EST-DB.

C- Les différents projets scientifiques impliqués dans ce projet. (Copie de 2007 et nouveaux projets).

- **Programme 1 : Symbioses actinorhiziennes (Equipe Rhizogenèse IRD, UMR DIA-PC ; Equipe Ecologie microbienne, CNRS/Univ Lyon 1, UMR 5557)**

Les travaux développés par l'équipe IRD Rhizogenèse symbiotique ont pour objectif de comprendre les mécanismes moléculaires et cellulaires qui aboutissent à la mise en place et au développement des racines symbiotiques des arbres tropicaux de la famille des Casuarinacées (Filao). Les Casuarinacées peuvent former des nodules racinaires en symbiose avec une bactérie filamenteuse du sol, *Frankia*. Les Casuarinacées appartiennent au groupe des plantes actinorhiziennes qui représente après les Légumineuses le deuxième groupe de plantes fixatrices d'azote. Les arbres tropicaux de la famille des Casuarinacées jouent un rôle environnemental essentiel, notamment pour les pays du Sud. Ces plantes possèdent une croissance rapide, sont bien adaptées à la sécheresse et sont capables de coloniser des sols pauvres.

En 2002, une étude comparative du transcriptome des racines et des nodules de *C. glauca* a été entreprise dans le cadre du GENOPOLE Montpelliérain et le séquençage de transcrits à partir de deux banques d'ADNc de *Casuarina* (racine et nodules) a permis l'obtention d'environ 3000 séquences. L'analyse bio-informatique sur la plateforme IRD a permis la validation des séquences et la création de la première base de données génomique pour les plantes actinorhiziennes.

Nous développons actuellement un projet visant à comparer plusieurs espèces de plantes actinorhiziennes ayant des caractéristiques différentes en terme de mode d'infection et de développement nodulaire. L'obtention en 2007 d'un projet Genoscope en association avec l'équipe de P. Normand (CNRS/Université Lyon I), va nous permettre le séquençage de 50 000 ESTs à partir de deux espèces actinorhiziennes (25 000 séquences pour *C. glauca* et 25 000 pour l'aulne, *Alnus glutinosa*) dont le traitement bioinformatique sera réalisé à l'IRD. Une analyse comparative des séquences produites devrait permettre l'identification *in silico* de gènes communs aux espèces actinorhiziennes activés lors de la mise en place de la symbiose. Par ailleurs, des études phylogénétiques récentes suggèrent une origine commune pour les différents types de symbioses fixatrices d'azote. La comparaison des séquences obtenues à celles disponibles pour les légumineuses devrait permettre de dégager des mécanismes moléculaires communs aux deux types de symbioses et ainsi de mieux comprendre les facteurs impliqués dans la mise en place des nodules racinaires fixateurs d'azote. Ce projet permettra la mise en place d'un réseau international qui contribuera au développement des ressources génomiques des plantes actinorhiziennes.

La réalisation de ce programme de recherche est conditionné par l'utilisation de la plateforme bio-informatique IRD qui nécessite (1) des optimisations pour permettre à nos partenaires un accès aux données génomiques et (2) l'ajout de nouvelles fonctions pour la réalisation des analyses de génomique comparative.

- **Programme 2 : *Aeschynomene* (Equipe Ecologie et Physiologie Moléculaire des Bradyrhizobium Photosynthétiques IRD, UMR LSTM)**

Le programme de recherche développé par l'Equipe IRD "Bradyrhizobium photosynthétique" a pour objectif d'identifier les acteurs moléculaires nécessaires à l'interaction entre les légumineuses tropicales du genre *Aeschynomene* et les Bradyrhizobiums photosynthétiques.

L'interaction Légumineuse-*Rhizobium* se caractérise par la spécificité avec laquelle elle s'établit. La reconnaissance mutuelle des deux partenaires est donc une étape clef de la mise en place de cette symbiose. Cette reconnaissance est rendue possible par une signalisation moléculaire complexe qui implique des lipo-chito-oligo-saccharides bactériens (LCOs ou facteurs NOD) reconnus par des kinases végétales particulières. Cet évènement de reconnaissance entre facteurs NOD bactérien et kinases végétales induit chez la légumineuse une cascade de signalisation complexe qui conduit à la formation d'un organe spécialisé, le nodule, au sein duquel la bactérie symbiote s'internalise. Cette suite d'événements a été mise en évidence chez tous les *Rhizobium* caractérisés jusqu'à maintenant, mais l'universalité de ce paradigme a été très récemment remise en question par notre étude du couple *Bradyrhizobium* photosynthétiques-*Aeschynomene*. En effet, l'examen des séquences génomiques de deux souches de *Bradyrhizobium* photosynthétiques (ORS278 et BTAi1) n'a pas permis de détecter la présence des gènes *nod* communs chez ces bactéries ; ceci démontre que les facteurs Nod ne sont pas requis pour induire l'organogénèse nodulaire chez *Aeschynomene* (Giraud *et al*. Science, 2007).

La dispense de facteurs Nod pour l'établissement du couple symbiotique Aeschynomene-*Bradyrhizobium* soulève de nombreuses questions : quelles sont la ou les molécules signal induisant l'organogénèse nodulaire ? Peut-on faire un parallèle entre les mécanismes moléculaires utilisés par les bradyrhizobia photosynthétiques et ceux mis en œuvre par d'autres organismes symbiotiques tels que *Frankia* qui induit la nodulation chez des non-légumineuses ? Pour répondre à cette dernière question, nous recherchons dans cette interaction originale, la présence d'éléments connus (récepteurs, facteurs de régulations…) de voies de signalisation décrites dans le cas de la symbiose fixatrice d'azote mais nous souhaitons également développer une approche sans *a priori* pour l'identification d'acteurs moléculaires originaux.

Dans cette optique, nous avons obtenu le soutien du Génoscope pour le séquençage de 56 000 ESTs à partir de deux espèces d'Aeschynomene dont le traitement bioinformatique sera réalisé à l'IRD. Ces deux espèces d'Aeschynomene appartiennent à deux groupes d'inoculation croisée distincts et correspondent à deux processus d'infection : l'un dépendant des facteurs Nod, l'autre indépendant des facteurs Nod. L'analyse comparative des séquences produites à partir de ces deux espèces devrait donc permettre de mettre en exergue les gènes spécifiques du processus symbiotique Nod indépendant. Ce projet s'inscrit dans un programme plus large visant à identifier les signaux « non-Nod » impliqués dans notre modèle mais également dans les symbioses actinorhiziennes et regroupe les équipes de P. Normand (CNRS/Université Lyon I), de D. Bogusz (Equipe Rhizogenèse IRD, UMR DIA-PC) et la nôtre.

- **Programme 3 : *Coffea (*Equipe génomique et qualit é du café, IRD** , **UMR DIA-PC*)***

L'équipe génomique et qualité du café s'engage dans un programme en génomique comparative au sein des *Rubiaceae* et entre *Rubiaceae* et *Solanaceae*, sans pour autant négliger la comparaison avec *Arabidopsis*.

On dispose de plusieurs milliers de séquences EST café produits par notre laboratoire ou par d'autres membres du réseau international génomique caféier (ICGN). Dans le cadre du réseau international RubiComp (*Rubiaceae* comparative) soutenu par l'IRD dans sa fonction d'Agence, il est prévu dans un avenir très proche de produire plusieurs milliers d'EST à partir de différents tissus et de banques soustraites de *Psychotria (Rubiaceae)*. La constitution d'une nouvelle banque BAC caféier est en projet et conduira dans un premier temps, au séquençage des extrémités des séquences BAC. Enfin, de très nombreuses données en génomique sous forme d'EST, de séquences de BAC et de séquençage de génome concernant la famille des *Solanaceae* (essentiellement la tomate) sont déjà disponibles.

Dans ce projet, nous nous intéressons à la comparaison de séquences et à l'identification de séquences orthologues intra et inter familles. Dans ce but, l'annotation homogène et cohérente des différentes banques permettra des recherches simplifiées de nouvelles séquences et constituera un système de référence. Les travaux en génomique comparative *via* la cartographie comparée (macro-synténie) entrepris dans notre équipe pourront être affinés à partir de l'analyse de séquences de BAC (micro-synténie).

- **Programme 4 : Triticées et *Medicago* (Equipe Diversité INRA, UMR DIA-PC)**

Les travaux de notre équipe ont pour objet l'analyse de la diversité des plantes cultivées et des formes sauvages apparentées et la compréhension des mécanismes évolutifs qui expliquent les patrons de diversité observés.

Dans ce contexte, nous avons engagé depuis quelques années, l'analyse du polymorphisme de séquence présent au sein de deux groupes taxonomiques importants pour l'amélioration des plantes : la sous famille des triticées qui contient les formes cultivées de blé et leurs espèces progénitrices et le genre *Medicago*, qui contient la principale espèce modèle pour les légumineuses, *M. truncatula* ainsi que la luzerne cultivée. Ces travaux s'inscrivent dans le cadre de l'analyse (i) de l'impact du processus de domestication sur la diversité des plantes cultivées et (ii) de la recherche de trace d'effets sélectifs dans le polymorphisme de

séquence. Dans les deux cas, l'analyse du polymorphisme de séquence concerne deux échelles taxonomiques: l'échelle intra-spécifique et l'échelle inter-spécifique, et plusieurs dizaines de fragments génomiques. Dès l'acquisition des premières données sur les Triticées, une interface WEB permettant le stockage et l'organisation des données de séquence a été mise en place (à travers l'encadrement de stages d'Informatique) : base de données Tritipol. Cette interface a aujourd'hui été clonée pour les données de séquence du projet *Medicago* : base de données Eagle ; un troisième « clonage » est en cours pour accueillir des données de séquence obtenue chez la Vigne (*Vitis vinifera* et formes sauvages apparentées) dans le cadre d'un projet similaire aux deux projets ci-dessus. Ces bases de données ont été conçues pour permettre aux différents partenaires des projets de récupérer les séquences et l'ensemble des informations relatives à l'origine de la donnée (extraction d'ADN, protocole d'amplification, informations sur les amorces, …). Nos premières analyses de données montrent aujourd'hui que ces interfaces pourraient être largement améliorées à travers : l'ajout d'une interface permettant le calcul de différentes statistiques résumant le polymorphisme et la mise en place de liens avec d'autres bases de données.

Chez *Medicago truncatula*, nous projetons de densifier l'analyse du polymorphisme de séquence le long d'un bras chromosomique (chromosome 5) afin d'accéder à des mesures de déséquilibre de liaison. Le séquençage des régions riches en gènes de ce chromosome est en cours (CNS Evry) et les données partiellement disponibles. Dans ce contexte, nous allons être amenés à définir des fragments génomiques balisant le chromosome 5 et des amorces permettant l'amplification spécifique de ces fragments. Pour définir ces amorces (~400 fragments prévus), il serait intéressant de pouvoir mettre en place un outils bioinformatique de routine permettant une recherche automatisée et systématique de fragments génomiques vérifiant certains critères comme par exemple l'absence de zones répétées, l'unicité au sein du génome de la région considérée, et la recherche dans les bases de données EST d'homologie de séquences avec des gènes connus chez d'autres Légumineuses d'intérêt (Pois, Soja, Haricot, Lotier, ..).

- **Programme 5 : Café / cotonnier (Equipe Résistances IRD, UMR 186 RPB)**

Nos objectifs sont d'identifier et de comprendre les mécanismes cellulaires, moléculaires et génétiques mis en jeu dans la résistance des plantes aux parasites. Plus précisément, nos recherches se focalisent d'une part, sur l'identification et la caractérisation fonctionnelle de gènes impliqués dans la résistance et l'activation des réactions de défense, et, d'autre part, sur l'exploration de la diversité des mécanismes de résistance associés à différentes interactions plante/parasite.

Nos modèles d'étude sont:

- le caféier (*Coffea arabica*) attaqué par le champignon *Hemileia vastatrix*, et les nématodes du genre Meloidogyne,
- le cotonnier (*Gossypium hirsutum*) infecté par la bactérie *Xanthomonas campestris* pv *malvacearum*.

Les activités développées font appel à des approches de génomique fonctionnelle, les approches transcriptomiques étant privilégiées. Chez le caféier, nous avons développé des banques d'ADNc soustractives pour établir un catalogue des gènes exprimés lors des réponses de résistance du caféier aux parasites et plusieurs gènes spécifiquement exprimés dans la résistance ont été clonés. Cependant, environ 35% des ESTs obtenues dans le cadre de l'interaction du caféier à *M. exigua* n'ont pu être annotées, faute de similarité avec des séquences connues, et pourraient représenter des séquences spécifiques des interactions plante/nématodes. D'autres banques d'ADNc sont en cours de construction avec nos partenaires Brésiliens (Embrapa) associés à ce projet, et nécessiteront l'utilisation d'outils bioinformatiques automatisés pour l'annotation des séquences. Chez le cotonnier, des approches physiologiques ont permis d'identifier plusieurs enzymes essentielles intervenant dans les voies de signalisation de la résistance (lipoxygénase, peroxydase, lipase) et les gènes correspondants sont en cours d'analyse fonctionnelle. Les recherches s'orientent vers la caractérisation de facteurs de transcription de type AP2 impliqués dans la voie de signalisation dépendante du jasmonate.

Pour les deux plantes, plusieurs milliers d'ESTs sont maintenant disponibles dans GenBank, mais ne représentent pas encore l'intégralité du génome transcrit. L'apport de la génomique comparative est donc essentiel à l'identification de nouveaux gènes, et à la caractérisation de leur fonction. Ainsi, par exemple, comme déjà précisé dans le programme 2, l'intégration des ressources génomiques de la famille des *Solanaceae*, proche de celle des *Rubiaceae* dont fait partie le caféier, permettra sans aucun doute d'identifier des orthologues de gènes clés de la résistance des plantes aux parasites. En particulier, un gène de résistance aux nématodes a été cloné chez la tomate, alors qu'aucun gène n'est encore connu chez *A. thaliana*, faute de résistance aux nématodes chez cette espèce modèle. Par contre, on pourra s'appuyer sur les connaissances de la famille AP2 chez *A. thaliana* pour isoler leurs orthologues chez le cotonnier. Les ressources bioinformatiques qui seront développées à l'IRD faciliteront ces recherches et l'annotation des nouvelles séquences.

- **Programme 6 : Palmier à Huile (Equipe Arécacées, IRD , UR 192 Palmiers**)

Le palmier à huile (famille *Arecaceae*, ordre Arecales) est une monocotylédone pérenne cultivée en zone inter-tropicale qui constitue, depuis 2004, la première source d,huile végétale dans le monde. De par sa grande productivité, cette plante est un enjeu clé pour le développement de l,agriculture durable dans de

nombreux pays tropicaux mais également dans l,approvisionnement de biocarburants sources d,énergies renouvelables au niveau mondial.

L'équipe Arécacées s,intéresse à différents aspects de la biologie du développement de cette plante dont des connaissances approfondies sont nécessaires pour pouvoir mettre à la disposition des planteurs un matériel végétal performant. Plus particulièrement, nous nous intéressons à la floraison (détermination de la structure florale, anomalies homéotiques de type épigénétique), à la fructification et à la formation de l,embryon (embryogenèses zygotique et somatique).

Afin d,étudier les processus de régulation sous-jacents à ces différents aspects du développement reproducteur, nous poursuivons, depuis plusieurs années, une approche de type transcriptomique. Ceci implique la constitution d,une collection d,étiquettes de séquence d,ADNc (EST) et leur utilisation pour effectuer des analyses d,expression différentielle à haut débit (expériences de type macroarray/microarray). A l,heure actuelle, la collection non redondante de séquences EST s'élève à plus de 9 000 séquences et continue de grandir. La collection de clones EST provient de plusieurs organes différents de la plante (inflorescence, pousses feuillées, embryons somatique et zygotique).

Dans le cadre d,une nouvelle collaboration initiée avec la Thaïlande, les ressources génomiques disponibles pour le palmier à huile vont être largement augmentées en 2008. Ces ressources, dont le traitement bioinformatique sera réalisé à l,IRD, nécessiteront l,utilisation d,outils bioinformatiques automatisés pour l,annotation des séquences. Elles serviront de base pour (1) le développement de la première puce à oligonucléotides de palmier à huile qui sera utilisée dans le cade d,un projet focalisant sur le développement et la maturation du fruit (bourse de thèse RTRA) et (2) la recherche de marqueurs SRR en collaboration avec le Cirad.

.

- **Programme 7: Hevea (Equipe Hévéa, IRD-Mahidol University, UR060/Clifa)**

Hevea brasiliensis, est la seule espèce végétale cultivée (zone tropicale humide) pour la production de latex, duquel est tiré le caoutchouc naturel.

Le programme « Recherche de marqueurs moléculaires du stress et de gènes candidats liés à la production du latex chez *Hevea brasiliensis* », menée par l'équipe Franco– Thaïe (IRD-Mahidol University) est basé sur l'analyse de l'expression différentielle de gènes dans la latex et le phloème (écorce interne) d'*Hevea*. L'étude porte sur des arbres de clones à haut et bas potentiel de production, soumis ou non à stress abiotiques (anthropiques ou environnementaux) conduisant à une surproduction transitoire (agents stimulants), ou au contraire à la cessation définitive de la production du latex (syndrome des « encoches sèches » ou de la « nécrose du phloème »).

L'étude est basée sur la construction et l'analyse de banques soustraites (SSH) d'ADNc de latex ou d'écorce interne des différents phénotypes étudiés. Six banques SSH ont déjà été construites en 2005 puis fin 2006, desquelles en tout environ 7.000 EST ont été séquencés. Quatre nouvelles banques SSH seront élaborées en 2007, avec un séquençage prévu d'environ 4600 nouvelles EST. D'autres programmes de séquençage à partir de nos banques d'ADNc pleine longueur, sont prévus dans un futur proche.

L'analyse bioinformatique des ces banques d'EST, au moyen du pipeline EST-DB de l'IRD-Montpellier, permettra le tri d'unigènes et l'élaboration, dans un premier temps, de filtres macroarray, puis à terme de microarrays. Ces futurs outils serviront au diagnostic pour l'optimisation de l'exploitation en plantation, et pour la sélection précoce de nouveaux clones performants, dans le cadre des programmes d'amélioration de l'hévéa, mis en œuvre au sein des instituts spécialisés des différents pays de la zone tropical humide, producteurs de caoutchouc naturel.

Ce programme nécessite l'utilisation de gros moyens de calcul et de fortes compétences en matière de bioinformatique et statistique. Dans le cadre de ce programme de recherche formation sur l'hévéa, deux chercheurs Thais, l'un de l'Université de Mahidol et l'autre de l'Institut BIOTEC (Bangkok), suivent une formation (2006-2008) en Mastère de Bioinformatique à l'UM2, avec stage pratique à l'IRD-montpellier sous la responsabilité de Christine Tranchant. Outre l'aide qu'il procurera au programme de recherche « Hevea », ce programme de formation, cofinancé par le MAE et le Ministère des Universités Thaïlandais, a pour but, à terme, d'initier un réseau d'agro-bioinformatique Thaïlandais, en coopération avec l'équipe de bio-informatique/GeneTrop de l'IRD-Montpellier. A cet effet, la plateforme bio-informatique est d'ores et déjà accessible par nos partenaires thailandais via le web, notamment l'application EST-db.

D- Liste des publications

Les publications soulignées sont celles ayant un rapport direct avec EST-db

Adam H, Jouannic S, Morcillo F, Verdeil JL, Duval Y, Tregear JW. (2007) Determination of flower structure in *Elaeis Guineensis*: do palms use the same homeotic genes as other species? *Annals of Botany*, 100:1-12

Adam H, Jouannic S, Orieux Y, Morcillo F, Richaud F, Duval Y, Tregear JW. (2007) Functional characterization of MADS box genes involved in the determination of oil palm flower structure. *J Ex Bot*, 1-15

BUSTAMANTE-PORRAS, J., CAMPA, C. PONCET, V., NOIROT, M., LEROY, **T.,** HAMON, S., de KOCHKO, A. (2007): Molecular Characterization of an Ethylene Receptor gene (CcETR1) in coffee trees. Its relationship

with fruit development and caffeine content. Mol. Genet. Geno. 277: 701–712

CAMPA, C., RAKOTOMALALA, J.J., de KOCHKO, A., HAMON, S. (2007): Chlorogenic Acids. Diversity in green beans of wild coffee species. Advances in Plant Physiology Accepted

Charoenwut c, P. Kongsawadworakul, J.P. Pichaut, D. Nandris, U. Sookmark, C. Tranchant, J. Narangajavana and H. Chrestin (2007)- Cloning and Characterization of Specific Molecular Markers of Rubber Tree Trunk Phloem Necrosis. In: Proc. IRRDB Int. rubb. Conf., 12-14 November 2007, Siem Reap, Cambodia,

Chatsapsin S, U. Sookmark, P. Kongsawadworakul, C. Tranchant and H. Chrestin (2007) - Differential expression of some ASR gene isoforms in the latex and bark of rubber tree. Effects of Ethrel stimulation. In: Proc. IRRDB Int. rubb. Conf., 12-14 November 2007, Siem Reap, Cambodia.

F. Cartieaux, C. Contesto, A. Gallou, G. Desbrosses, L. Taconnat, J-P. Renou and B. Touraine. (2007). Simultaneous interaction of *Arabidopsis thaliana* with *Bradyrhizobium* sp. ORS278 and *Pseudomonas syringae* pv *tomato* DC3000 leads to complex transcriptome changes. MPMI (accepté).

Fernandez D., Ramiro D., Petitot A.-S. and Maluf M. 2006. Phylogenetic analysis of the WRKY transcription factors gene superfamily in coffee plants. Proceedings of the 21st International Conference on Coffee Science, ASIC, Montpellier.

Fernandez D., Santos P., Agostini C., Bon M.-C., Petitot A.-S., Silva M. C., Guerra-Guimarães L., Ribeiro A., Argout X. and Nicole M. 2004. Coffee (*Coffea arabica* L.) genes early expressed during infection by the rust fungus (*Hemileia vastatrix*). Molecular Plant Pathology ,5, 527-536.

Ganesh D., Petitot A.-S., Silva M., Alary R., Lecouls A.C. and Fernandez D. 2006. Monitoring of the early molecular resistance responses of coffee (Coffea arabica L.) to the rust fungus (Hemileia vastatrix) using real-time quantitative RT-PCR. Plant Science, 170:1045-1051.

Giraud E., Moulin L., D. Vallenet, V. Barbe, E. Cytryn, J-C. Avarre, M. Jaubert, D. Simon, F. Cartieaux, Y. Prin, G. Bena, L. Hannibal, J. Fardoux, M. Kojadinovic, L. Vuillet, A. Lajus, S. Cruveiller, Z. Rouy, S. Mangenot, B. Segurens, C. Dossat, W. L. Franck, W-S. Chang, E. Saunders, D. Bruce, P. Richardson, P. Normand, B. Dreyfus, D. Pignol, G. Stacey, D. Emerich, A. Vermeglio, C. Medigue, And M. Sadowsky. (2007). Legumes Symbioses: Absence of *Nod* Genes in Photosynthetic Bradyrhizobia. Science 316: 1307-1312

Hocher V., Auguy F., Argout X., Laplaze L., Franche C., and Bogusz D. Expressed sequence tag analysis in *Casuarina glauca* actinorhizal nodule and root. New Phytologist. 2006 169:681-688.

Jaubert M, Lavergne L, Fardoux J, Hannibal L, Vuillet L, Adriano J-M, Bouyer P, Pignol D, Giraud E*, Verméglio A*. (2007) A singular bacteriophytochrome acquired by lateral gene transfer. J. Biol. Chem. 282:7320-8 *Co-senior authors

Jouannic S, Collin M, Vidal B, Verdeil JL, Tregear JW. (2007) A class I KNOX gene from the palm species Elaeis guineensis (Arecaceae) is associated with meristem function and a distinct mode of leaf dissection. *New Phytologist*. 174:551-568

Jouannic, S., Argout, X., Lechauve, F., Fizames, C., Borgel, A., Morcillo, F., Aberlenc-Bertossi, F., Duval, Y., and Tregear, J. (2005). Analysis of expressed sequence tags from oil palm (*Elaeis guineensis*). FEBS Letters 579, 2709-2714.

Kojadinovic, M., Laugraud, A., Vuillet, L. Fardoux, J., Hannibal, L., Adriano, J.M., Bouyer, P., Giraud, E., Vermeglio, A. (2007). Dual role for a bacteriophytochrome in the bioenergetic control of rhodopsdeudomonas palustris: enhancement of photosystem synthesis and limitation of respiration. BBA Section: BBA – Bioenergetics (accepté).

Konan KE, Durand-Gasselin T, Kouadio YJ, Niamké AC, Dumet D, Duval Y, Rival A & Engelmann F (2007) Field development of oil palms (*Elæis guineensis* Jacq.) originating from cryopreserved Stabilized Polyembryonic Cultures (SPCs). *Cryo Letters 28/5:377-386*

Kongsawadworakul P., Sookmark U., Nandris D. and H. Chrestin (2005) - Cyanide metabolism and molecular approach of rubber trunk phloem necrosis: Present and prospects (oral communication). In: Proc. Int. Hevea workshop on tapping panel dryness. Kerala, India, November 2005.

LAFARGUE, A., de KOCHKO, A., DUSSERT, S. (2007): Development of solid-phase extraction and methylation procedures to analyse free fatty acids in lipid-rich seeds. Plant Physiol. Biochem. 45 (3-4): 250-257

Lecouls A.-C., Petitot A.-S. and Fernandez D. 2006. Early expressed genes in the coffee resistance response to root-knot nematodes (Meloidogyne sp.) infection. Proceedings of the 21st International Conference on Coffee Science, ASIC, Montpellier.

Lucas M., Godin C., Jay-Allemand C. and Laplaze L. Auxin fluxes in the root apex co-regulate gravitropism and lateral root initiation. J. Exp. Bot. (2007). In press.

MAHESH, V., ULLMANN, P., NOIROT, M., HAMON, S., de KOCHKO, A., WERCK-REICHHART, D., CAMPA, C., (2007): CYP98A-C1 and CYP98A-C2 encode p-coumaroyl 3’-hydroxylases mediating chlorogenic acid biosynthesis in coffee trees. Plant Mol. Biol. 64: 145-159

Morcillo F, Gallard A, Pillot M, Jouannic S, Aberlenc-Bertossi F, Collin M, Verdeil JL, Tregear JW (2007) *EgAP2-1*, an *AINTEGUMENTA-like* (*AIL*) gene expressed in meristematic and proliferating tissues of embryos in oil palm. *Planta*. 226(6):1353-62.

N'DIAYE, A., M. NOIROT, S. HAMON, AND V. PONCET. 2007. Genetic basis of species differentiation between Coffea liberica and C. canephora: analysis of an interspecific cross. Genetic Resources Crop Evolution 54: 1011-1021.

Obertello M., Wall L., Laplaze L., Nicole M., Auguy F., Gherbi H., Bogusz D. and Franche C. Functional analysis of the metallothionein gene CgMT1 isolated from the actinorhizal tree Casuarina glauca. Mol. Plant-Microbe Interact. (2007). 20: 1231-1240.

Péret B., Svistoonoff S., Lahouze B., Auguy F., Santi C., Doumas P. and Laplaze L. A role for auxin during actinorhizal symbioses formation ? Plant Signal Behav. (2008). In press.

Péret B., Swarup R., Jansen L., Devos G., Auguy F., Collin M., Santi C., Hocher V., Franche C., Bogusz D., Bennett M. and Laplaze L. Auxin influx activity is associated with Frankia infection during actinorhizal nodule formation in Casuarina glauca. Plant Physiol. (2007). 144: 1852-1862. (Download PDF)

Petitot A.-S., Lecouls A.C. and Fernandez D. 2007. Sub-genomic origin and regulation patterns of a duplicated WRKY gene in the allotetraploid species Coffea arabica. Tree Genetics and Genomes, DOI 10.1007/s11295-007-0117-x.

PONCET, V., DUFOUR, M., HAMON, P., HAMON, S., de KOCHKO, A., LEROY, T. (2007) : Development of genomic microsatellite markers for *Coffea* genus and their potential use for endangered wild species. Genome. Accepted

Poncet, V., Rondeau, M., Tranchant, C., Cayrel, A., Hamon, S., de Kochko, A., Hamon, P. (2006). SSR mining in coffee tree est databases: potential use of EST-SSRs as marker across *Coffea* genus. Mol. Genet. Geno. 276, no. 5, pp. 436-449.

Rotchanapreeda T, P. Kongsawadworakul, U. Sookmark, C. Tranchant and H. Chrestin (2007) - Ethylene induces Lipoxygenase Genes (jasmonate pathway) early upregulation in the inner bark tissues of Rubber Tree. In: Proc. IRRDB Int. rubb. Conf., 12-14 November 2007, Siem Reap, Cambodia.

SALMONA, J., DUSSERT, S., DESCROIX, F., de KOCHKO, A., BERTRAND, B., JOËT, T., (2007) : Deciphering transcriptional networks that govern *Coffea arabica* seed development using combined cDNA array and real-time rt-pcr approaches. Pl. Mol. Biol. Accepted

Silva M.C., Várzea V., Guerra-Guimarães L., Azinheira H.G., Fernandez D., Petitot A.-S., Bertrand B., Lashermes P., Nicole M. 2006. Coffee resistance to the main diseases: leaf rust and coffee berry disease (CBD). Brazilian Journal of Plant Physiology,18:119-147.

Sy M.O., Hocher V., Gherbi H., Laplaze L., Auguy F. Bogusz D. and Franche C. The cell-cycle promoter cdc2aAt from Arabidopsis thaliana is induced in the lateral roots of the actinorhizal tree Allocasuarina verticillata during the early stages of the symbiotic interaction with Frankia. Physiol. Plant. (2007)

Verdeil JL., Niemena N., Alemanno L., Tranbarger Timothy John. (2007) Pluripotent versus Totipotent plant stem cells: dependence versus autonomy? *Trends In Plant Science*, Vol.12 N°6:245-252

Vuillet, L., Kojadinovic, M., Zappa, S., Jaubert, M., Adriano, J.M., Fardoux, J., Hannibal, L., Pignol, D., Verméglio, A. and Giraud, E. (2007) Evolution of a bacteriophytochrome from light to redox sensor. *The EMBO Journal: 6*(14):3322-31

34. Description et analyse des besoins (Copie 2007)

Depuis quelques années, les différentes équipes IRD du domaine végétal développent plusieurs projets de génomiques. Chaque projet a généré une masse importante d'informations qui était impossible d'analyser et d'exploiter sans l'aide de la bio-informatique.

Dès 2001, ces équipes IRD Plante appartenant à différentes UMRs ont décidé de développer une application commune afin, d'une part, de mettre en place une chaîne de traitement permettant d'analyser les ESTs et, d'autre part, de créer une base de données (et le site web associé) destinés à gérer/mutualiser/mieux exploiter les informations générées par la chaîne de traitement. L'application EST-db a été développée au cours de 5 stages de master « Informatique Pour les Sciences » de l'Université de Montpellier. Elle est installé sur la plate-forme bio-informatique IRD dédié à la génomique végétale et est utilisée par les 4 UMRs du domaine végétal basées sur le centre IRD de Montpellier ainsi que par des partenaires et IRDiens expatriés (ex : Projet Hévéa, Thaïlande, Projet Café, Ile de la Réunion). Le pipeline a aussi été utilisé pour analyser des données d'autres UMRs telles que des ESTs issues de la souris. A l'heure actuelle, plus de 100000 ESTs ont été générées/analysées. Ce volume de données ne cesse d'augmenter et de nouvelles analyses sont demandées par les chercheurs. Il s'avère donc nécessaire de développer de nouvelles fonctionnalités sur l'outil EST-db.

Le projet SPIRALE que nous proposons a pour objectifs :

- **de finaliser le développement de l'application EST-db**(documentation du code, code

suffisament paramétrable pour que l'application soit facilement transférable sur une autre plate-forme bio-informatique, ajout de nouvelles fonctions qui donneront une plus value importante à l'application) en vue de la déclarer à l'Agence de Protection des Logiciels et de la distribuer aux autres URs de l'IRD et partenaires interessés par l'outil (2007)

- **de développer deux nouvelles fonctionnalités à EST-db**dédiés à l'annotaion automatique des séquences et à la réalisation d'analyse de génomique comparative.

  - Le module "Annotation Automatique"

    Ce module est un pré-requis indispensable à toute analyse de génomique comparative. Il s'agit de pourvoir classer les séquences ESTs annotées par le logiciel Blast en grande fonction selon la nomenclature Gene Ontology. Ce système utilisé par la communauté scientifique internationale est disponible mais nécessite une adaptation afin de s'inclure au pipeline EST-db existant.

  - Le module "Génomique comparative".

    Ce module permettra aux équipes plantes (IRD et partenaires) de réaliser des comparaisons entre les génomes des diverses plantes étudiées et/ou avec les génomes des plantes modèles. Ces analyses optimiseront l'identification de séquences orthologues entre différentes espèces et donc l'annotation des gènes identifiés. Ces comparaisons permettront de rechercher les relations existantes entre les gènes de différentes espèces (synténie) ainsi que les relations de ces gènes au sein d'un même génome.

Pour la réalisation du projet, une démarche classique de génie logicielle sera faite : phase d"analyse et de modélisation (interviews des acteurs, notation UML, cahier des charges, veille technologique, bibliographique), phase de développement, phase de test et de rédaction des documentations techniques, L'objectif final est l'obtention d'une application finalisée et générique permettant l'analyse des ESTs aussi bien issues de plantes que d'animaux. La déclaration à l'APL de l'application EST-db permettra sa distribution à toute équipe de recherche intéressée (IRD, extérieure, partenaires du Sud).

35. Description de l'existant (moyens – outils – compétences)

OU renvoyer à un document joint

A- Moyen

Tout développement sera réalisé sur la plate-forme bio-informatique dédié à la génomique végétale dont l'infrastructure est la suivante :

- 2 serveurs de calcul de production (DellTM PowerEdgeTM 6650 - 4 processeurs Xeon 2.7 Ghz & 8 Go de RAM)
- 1 serveur de développement
- Un système de stockage RAID Dell/EMC (1 To de stockage)
- Un serveur de fichier

B- Outil

Le pipeline et la base de données EST-db

Chaque projet génomique génère une masse importante d'informations sous la forme d'ESTs. Avant d'aboutir à son annotation, chaque EST doit subir une série de traitements nécessitant l'utilisation de différents logiciels. Compte tenu du volume important d'informations et de traitements, un pipeline est nécessaire pour automatiser l'analyse de chaque EST :

- A l'issue du séquençage des ESTs, les chromatogrammes générés sont analysés afin d'obtenir la séquence nucléïque.
- La séquence ensuite doit être analysée afin de masquer les bases de mauvaise qualité et celles appartenant au vecteur puis sélectionner uniquement les parties réellement informatives.
- Afin de supprimer la redondance au niveau des séquences et de déterminer l'agencement des séquences, une phase de contiguage est nécessaire.
- Puis, les séquences sont annotées.

L'étape finale est de comparer un pool d'ESTs avec celui produit sur une autre plante tropicale ou avec les données publiques.
Les résultats de chaque étape de ce traitement sont archivés dans une base de donnée et l'ensemble (pipeline – base de donnée) est consultable et utilisable au travers d'une interface Web.

Structure du pipeline

Le "pipeline" ou la chaîne de traitements est un programme perl qui va permettre de combiner

l'exécution de plusieurs logiciels, l'analyse des résultats générés et de réaliser d'autres fonctionnalités répondant à des critères propres au laboratoire. Les données brutes et les données générées sont ensuite stockées dans une base de données MySQL.

A l'issue du séquençage, la séquence d'ADN est représentée par un chromatogramme qui va être analysé par le pipeline. Le logiciel de « base calling » utilisé est **Phred** qui va permettre d'obtenir les séquences nucléiques des ESTs.

Puis, les résultats de Phred sont traités : les bases de mauvaises qualités sont masquées et les séquences de mauvaise qualité sont éliminées.

Les séquences appartenant au vecteur sont ensuite détectées à l'aide du logiciel **Vecscreen** puis elles sont masquées et supprimées. Les séquences de petite taille sont éliminées.

Chaque séquence d'EST représente un fragment d'un génome mais certaines d'entre elles peuvent être redondantes ou recouvrantes. Le contigage des séquences va permettre de réduire le nombre de séquences à annoter, d'obtenir des séquences plus longues et donc réaliser une annotation plus fiable. Ceci est réalisé par le logiciel **Stackpack**. A l'issue du contigage, les ESTs appartiennent ou non à un contig. L'étape suivante est l'annotation des séquences qui doit renseigner sur la fonction des protéines putatives éventuellement associées. Une des méthodes les plus sures pour la détermination des gènes est la comparaison de la séquence à analyser avec une banque de séquences. Il s'agit d'une approche par similitudes qui est réalisée par le programme d'alignement local Blast

Bases de données

-Swiss-Prot, est une base de données de séquences protéiques, qui possède un haut niveau d'annotations tel que la description des fonctions protéiques, les structures des domaines, les modifications post transcriptionnelles L'ensemble des données qui sont insérées dans Swiss Prot sont vérifiées manuellement par des « curateurs » qui rajoutent les informations dans la base, consultable sur le web ou sous forme de fichiers plats par FTP (http://us.expasy.org/sprot/).

-**Genbank** est une base de données de séquences nucléïques publiques regroupant 32549400 séquences en février 2004. (http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html)

-TIGR Rice Database, est une base d'annotation automatique (90% actuellement) de la séquence complète du génome du riz (IRGSP). La plupart des annotations ne sont pas vérifiées comme sur Swiss–Prot, même si le processus de curation manuel est en cours actuellement. Les données sont consultables via leur site web ou sous forme de fichiers XML par FTP
(http://www.tigr.org/tdb/e2k1/osa1/).

Logiciels

-**Phred**

Le programme Phred, développé en C, est un logiciel de « base calling » qui se base sur la méthode de Fournier pour lire les 4 courbes du chromatogramme. Il appelle une à une les bases, leur assigne une valeur de qualité et écrit les résultats dans des fichiers de sortie.
http://www.phrap.org/phredphrapconsed.html

-Vecscreen

Cet outil disponible sur NCBI permet d'identifier rapidement des segments de séquences nucléiques d'origine vectorielle. Il recherche la position du vecteur dans une séquence en utilisant le programme Blast automatiquement paramétré pour une détection optimale des contaminations.
http://www.ncbi.nlm.nih.gov/VecScreen/

-Stackpack

Ce logiciel possède un programme réalisant le clustering, l'assemblage de séquences présentant de courtes régions chevauchantes. Il regroupe 3 algorithmes :

- l'algorithme agglomératif *D2_cluster*, plus rapide que blast est utilisé pour le clutering initial; les séquences doivent être longues et seules les grandes similarités sont détectées.
- L'algorithme du programme *phrap* aligne rapidement toutes les séquences d'un cluster entre elles mais les informations sur la variation à l'intérieur même du cluster sont insuffisantes pour établir une séquence consensus.
- Le programme *craw* intervient dans l'ultime étape pour analyser l'alignement et déterminer la séquence consensus.

Ce logiciel est développé en python et toutes les informations manipulées par ce logiciel sont stockées dans une base de données Mysql.
http://www.sanbi.ac.za/Dbases.html

-Blast

BLAST (Basic Local Alignment Search Tool) est un programme de recherche de similarité développé au NCBI / Genbank. L'intérêt de l'algorithme est que sa conception est basée sur un modèle statistique. Celui-ci a été établi d'après les méthodes statistiques de Karlin et Altschul (1990 ; 1993) qui s'appliquent aux comparaisons de séquences sans insertion-délétion. L'unité fondamentale de BLAST est le HSP (High-scoring Segment Pair). Un HSP correspond à une région de similitude la plus longue possible entre deux séquences ayant un

score supérieur ou égal à un score seuil. Un deuxième score MSP (Maximal-scoring Segment Pair) a été défini comme étant le meilleur score obtenu parmi tous les couples possibles que peuvent produire deux séquences. Les méthodes statistiques de Karlin et Altschul sont appliquées pour déterminer la signification biologique des MSPs et par extrapolation la signification des scores HSPs obtenus lors de la comparaison.
http://www.ncbi.nlm.nih.gov/Tools/

C- Compétences

La plate-forme bio-informatique est administrée par le SIL de Montpellier et le service bio-informatique. Ce service se compose d'un ingénieur d'étude en bio-informatique et de personnels contractuels (stages, vacation, CDD).

36. Difficultés rencontrées jusqu'à présent :

6. Description du projet – SEULEMENT SI « développement d'application IS » (méthodes, solutions, et moyens)

Cette partie (question 37 à 96) concerne les **nouveaux projets finalisés de développement d'application IS ainsi que les demandes de continuums portant également le développement d'application IS.**

La demande peut-être être **accompagnée de tous documents utiles** :

présentation du projet global ou descriptif du projet, rapport de phases préliminaires, étude de faisabilité, dossier d'expression des besoins ou ***cahier des charges****, devis détaillé…*

37. Nom de votre outil

EST-db

38. Si votre outil existe déjà, quel est l'URL du site internet ou des documents qui le décrivent? Ou, si l'outil a été décrit dans un article, fournir les références

L'application est disponible en intranet sur http://bioinfo/estdb (consultation des données et lancement du pipeline). Elle est également ouverte pour des partenaires ou IRD expatriés en consultation sur http://www.mpl.ird.fr/bioinfo (site de la plateforme bioinformatique comportant un lien vers EST-db; accès soumis à l'authentification) après demande d'accès et signature d'une charte.

Cf : document joint documentation scientifique

🗀 Innovation :

39. Ecrire 3 scénarios qui illustrent comment votre outil sera ou a été utilisé dans votre communauté scientifique ou domaine d'activités

1 – Utilisation pour l'analyse d'EST : blast, annotation fonctionnelle Gene Ontology. De nouveaux projets sont concernés (Cf Programme 2)
2 – Utilisation pour la génomique comparative et les études de phylogénie (alignements des génomes ou de banques EST de différents organismes)
3 – Exportation de séquences EST sous format fasta pour des traîtements ultérieurs (e.g., recherche de microsatellites par le SSR pipeline mis en place aussi à l'IRD de Montpellier ; design de puces à ADN…)
4-Utilisation en consultation à distance par partenaires ou IRD expatriés (P. Normand, Lyon; H. Chrestin, tahïlande; T Joet, La réunion) En projet : transfert EST-db sur Thaïlande (H. Chrestin, C. Tranchant) et sur Nouméa (V. Poncet)

40. Décrire, en un paragraphe, les innovations de votre projet pour votre communauté scientifique

Les équipes IRD concernées par ce projet utilisent déjà en routine l'application EST-db pour traiter leurs données de génomique. Ce projet permettra de réaliser de nouveaux types d'analyses indispensables comprendre la biodiversité et rechercher de nouveaux gènes d'intérêts chez les plantes tropicales. EST-db est un outil incontournable pour la réalisation de nos programmes de recherche et pour leur valorisation sous forme de publication. Ce projet SPIRALE permettra aussi l'optimisation d'EST-db afin de rendre cette application transférable et utilisable pour les équipes IRD expatriées et les partenaires extérieurs.

41. Existent-ils d'autres outils similaires au vôtre ? Si c'est le cas, lister ces outils et décrire les avantages de votre outil par rapport aux autres

Lors de la mise place du système EST-db (2001), il n'existait aucun pipeline générique d'analyse d'ESTs excepté des applications payantes et relativement chères. Nous avons donc opté pour développer notre propre programme qui enchaîne plusieurs analyses réalisés par des logiciels bio-informatiques gratuits aux algorithmes robustes. A l'heure actuelle, d'autres laboratoires ont développé leur propre pipeline mais aucun ne distribue leur outil. Notre objectif est justement de rendre notre application suffisamment convivial pour le distribuer gratuitement après signature d'une charte. Nous avons déjà plusieurs demandes d'équipes IRD ou extérieures. Les nouvelles fonctionnalités que nous avons développé ou souhaitons développer n'existent pas en tant que telles. Il s'agira de rechercher et d'adapter des outils existants afin qu'ils s'intègrent dans le système EST-DB et qu'ils répondent au mieux à nos besoins. L'objectif, une fois de plus, est de pouvoir le distribuer afin d'éviter que le travail soit refait.

42. Si vous proposez des améliorations à un outil existant, combien d'utilisateurs ont déjà téléchargés ou obtenus une copie de la version actuelle ?

Aucun, puisque l'objectif de ce projet est de rendre cet outils transférable (documentation, interface WEB facile, déclaration à l'APL)

43. Le projet proposé est-il basé sur de nouvelles conclusions scientifiques ou méthodes innovantes ? Si c'est le cas, décrire les fondements et lister les références les plus pertinentes.

Le projet est basé sur 2 principes :

- Ajout de nouvelles focntionnalités (Gene Ontology et Génomqiue Comparative) afin d'avoir un outils de traitement des ESTs complets. Ces modules sont nécessaire pour répondre aux éxigences des publications internationales.

- Finalisation de la documentation technique d'EST-dbafin de rendre transférable au sein et hors de l'IRD de façon gratuite. Ce point est essentiel dans le contexte IRD puisque différents partenaires du Sud sont demandeurs.

Calendrier, budget et risques

44. Calendrier du projet montrant les tâches clés et les dates d'échéances

Les dates d'échéance sont données à titre indicatif, car cela dépendra d'une part de l'acceptation du projet et de la mise à disponibilité des crédits.

A- Réalisation des actions prévues en 2007 et reportées sur 2008 (1 mois, fin janvier 2008)

- **Finalisation du développement d'EST-db :** documentation technique à mettre à jour après les dernières modifications
- **Amélioration d'EST-db :** lancement du blast 2 fois / an
- **Génomique comparative :**. Phase de développement

B- Réalisation des actions 2008

- **Amélioration du module « Project _ GO annotation » (3 semaines, fin février 2008)**
  - Annotation des contigs
- **Améliorer la version actuelle de l'application EST-db:** ajout de pages d'aide et/ou d'un manuel utilisateur **(2 semaines, début mars 2008)**
- **Amélioration de l'interface de lancement du pipeline d'analyse d'ESTs :** lancement du pipeline à partir des séquences (format fasta) et pas uniquement des chromatogrammes. **(3 semaines, fin mars**

2008)

- **Améliorations du pipeline (3 semaines, fin avril 2008)**
  - Possibilité de lancer plusieurs pipelines en parallèle
  - Sauvegarde dans la base de données EST-db des paramètres de lancement du pipeline qui sont indiqués dans l'interface de lancement de pipeline pour un projet d'ESTs.

- **Amélioration de la structure de la base de données EST-db (2 semaines, début mai(3 semaines, fin mai 2008)gfy 2008)**
  - Ajout d'attributs dans la base de données
  - Traduction en anglais les noms des tables et des attributs qu'elles contiennent (en vue d'une diffusion large).

- **Amélioration des interfaces de consultation (3 semaines, fin mai 2008)**
  - Modification de la page « All statistics » (affichage des paramètres de lancement du pipeline, affichage du nombre de séquences classées dans chaque catégorie GO, lien vers les séquences no hit;
  - Affichage du nombre d'EST/ contig et du pourcentage de singletons: banque
  - Modification de la page « Search by keyword »: (recherche par mots clé sur les blast et sur les terme GO);
  - Création d'une page « export ontology »: « Exporter toute l'ontologie avec le nombre d ESTs et le nom des ESTs comme sur GOBLET et utilisation par exemple d'AMIGO pour visualiser le graphe.
  - Ajout d'une fenêtre permettant à chaque consultant d'inclure des commentaires sur une séquence particulière (singleton et/ ou contig).

- **Nouvelles fonctions (1 mois, fin juin 2008)**
  - Un module « Publish EST » qui permettrait de formater les informations relatives à une sélection d'ESTs avec le format requis pour une soumission à une base de données publiques.
  - Module de recherche des Open Reading Frame (ORF)
  - Module permettant le calcul du codon usage et du % en G-C des séquences

45. Eventuellement, budget détaillé montrant les coûts des tâches clés, des différents modules ou phases.

(*Les informations apportées doivent être cohérentes avec celles précisées à la question 18.*)

Prestation de services : 20 000 euros, le détail n'est pas disponible ; Nous n'avons pas encore fait l'étude détaillée avec le prestataire.

Matériel : 10 000 euros (financement demandé inter UMR)

46. Si vous demandez des fonds pour des activités autres que du « développement logiciel », pourquoi ces activités sont-elles essentielles à l'accomplissement de votre projet ?

47. Quel sont les risques encourus si votre projet ne peut être finalisé à échéance et dans le budget prévu ? Comment comptez-vous pallier à ces risques ?

SI le projet ne peut pas être finalisé à l'échéance (toutes les améliorations de l'application EST-dbne sont pas implémentées), nous aurons néanmoins un produit utilisable par les équipes de l'IRD de Montpellier mais sa diffusion immédiate (e.g. transfert aux équipes sur Nouméa) ne sera pas possible. Pour pallier à ce risque, un cahier de charges détaillé sera rédigé, définissant l'ordre et la priorité des taches à effectuer ainsi que les taches critiques. Un suivi très rigoureux de l'avancement du projet sera mis en place, avec des réunions régulières avec le prestataire afin de faire un bilan et, en cas de retard, analyser ses causes et corriger éventuellement le planning, avec une prise en charge possible d'une certaine partie du travail par la responsable de la plateforme bioinformatique si nécessaire.

48. Si vous demandez un soutien d'un an, accepteriez-vous de recevoir les crédits l'année prochaine plutôt que cette année ?

49. Si cette demande concerne la phase 1 d'un projet prévu sur 2 ans, pouvez-vous réaliser le projet en entier sur une année si vous obtenez les crédits en une seule fois ? Comment cela

impacterait-il votre projet ?

🗀 Architecture de l'outil

50. Décrire l'architecture envisagée pour votre outil. Identifier les composants clés de l'application et décrire comment ils interagissent.

(*un schéma peut être appréciable*)

Le pipeline et la base de données EST-db

Chaque projet génomique génère une masse importante d'informations sous la forme d'ESTs. Avant d'aboutir à son annotation, chaque EST doit subir une série de traitements nécessitant l'utilisation de différents logiciels. Compte tenu du volume important d'informations et de traitements, un pipeline est nécessaire pour automatiser l'analyse de chaque EST :

- A l'issue du séquençage des ESTs, les chromatogrammes générés sont analysés afin d'obtenir la séquence nucléïque.
- La séquence ensuite doit être analysée afin de masquer les bases de mauvaise qualité et celles appartenant au vecteur puis sélectionner uniquement les parties réellement informatives.
- Afin de supprimer la redondance au niveau des séquences et de déterminer l'agencement des séquences, une phase de contiguage est nécessaire.
- Puis, les séquences sont annotées (blast, GO).

L'étape finale est de comparer un pool d'ESTs avec celui produit sur une autre plante tropicale ou avec les données publiques (module génomique comparative).
Les résultats de chaque étape de ce traitement sont archivés dans une base de donnée et l'ensemble (pipeline – base de donnée) est consultable et utilisable au travers d'une interface Web.

Structure du pipeline (voir le schéma ci-dessous)

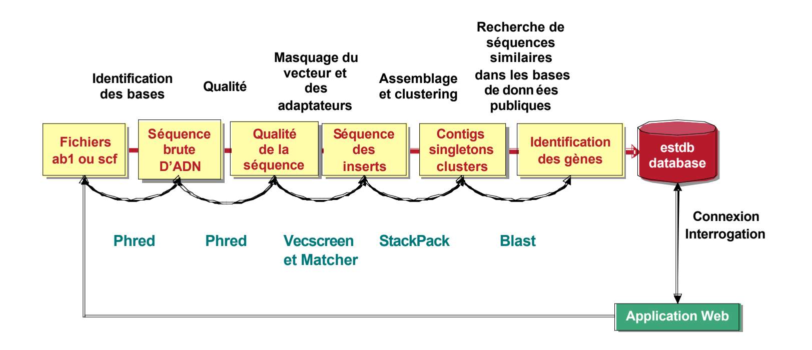
Le "pipeline" ou la chaîne de traitements est un programme perl qui va permettre de combiner l'exécution de plusieurs logiciels, l'analyse des résultats générés et de réaliser d'autres fonctionnalités répondant à des critères propres au laboratoire. Les données brutes et les données générées sont ensuite stockées dans une base de données MySQL.

A l'issue du séquençage, la séquence d'ADN est représentée par un chromatogramme qui va être analysé par le pipeline. Le logiciel de « base calling » utilisé est **Phred** qui va permettre d'obtenir les séquences nucléiques des ESTs.

Puis, les résultats de Phred sont traités : les bases de mauvaises qualités sont masquées et les séquences de mauvaise qualité sont éliminées.

Les séquences appartenant au vecteur sont ensuite détectées à l'aide du logiciel **Vecscreen** puis elles sont masquées et supprimées. Les séquences de petite taille sont éliminées.

Chaque séquence d'EST représente un fragment d'un génome mais certaines d'entre elles peuvent être redondantes ou recouvrantes. Le contigage des séquences va permettre de réduire le nombre de séquences à annoter, d'obtenir des séquences plus longues et donc réaliser une annotation plus fiable. Ceci est réalisé par le logiciel **Stackpack**. A l'issue du contigage, les ESTs appartiennent ou non à un contig. L'étape suivante est l'annotation des séquences qui doit renseigner sur la fonction des protéines putatives éventuellement associées. Une des méthodes les plus sures pour la détermination des gènes est la comparaison de la séquence à analyser avec une banque de séquences. Il s'agit d'une approche par similitudes qui est réalisée par le programme d'alignement local Blast



Shéma actuel du pipeline et de la base de donnée EST-db. Ce shéma ne prend pas en compte les modifications prévues dans ce projet.

Cf aussi la documentation technique où un schéma plus détaillé montrant l'architecture est présenté.

51. Lister les méthodes/référentiels d'analyses, de conception et de développement utilisés pour élaborer l'outil.

- Analyse et conception : notation UML (Unified Modeling Language)
- Développement : Perl, CGI ; XHTML, CSS, Javascript
- Système de gestion de bases de données MySQL

52. Lister les langages de programmations et les outils de développement envisagés. Préciser le type de syntaxe qui sera utilisée pour la documentation du code.

Langages de programmations et outils de développement utilisés :

- Langages de programmation : Perl, bioPerl, Python, XHTML, CSS, JavaScript
- Le logiciel libre Umbrello (Linux) est utilisé pour la modélisation UML.
- L'éditeur de texte Emacs sous Linux est utilisé pour l'écriture des scripts.
- L'accès à la base de données se fait via terminal ou via l'interface phpMyAdmin.

Documentation du code :

Tous les scripts sont soigneusement commentés. Pour la documentation du code, la syntaxe standard propre à chaque langage est utilisée.

Ainsi, pour les scripts Perl les commentaires ponctuels sont introduits par # Commentaire.

Description d'une fonction en Perl:

```
#begin
## Name :
## Actions :
## Parameters :
## Returns :
#end
```

Documentation du code JavaScript : les commentaires sont inclus dans /*Commentaire*/.

Pour un script/module, l'entête du fichier comportera sa description générale :

```
######################################################################
###                                                                ###
### GENERAL INFORMATIONS :                                         ###
###                                                                ###
### Name of the script :                                           ###
###                                                                ###
```

```
### Programming language :                                            ###
###                                                                   ###
### Authors :                                                         ###
###                                                                   ###
### Location :                                                        ###
###                                                                   ###
### Updated :                                                         ###
###                                                                   ###
### Script role :                                                     ###
###                                                                   ###
###                                                                   ###
### LINK WITH :                                                       ###
###                                                                   ###
###                                                                   ###
#####################################################################
###             HELP : bioinfo@mpl.ird.fr                             ###
#####################################################################
```

En vue de la déclaration de l'application à l'APL, tous les commentaires seront traduits en anglais.

53. Lister le matériel et les logiciels requis pour faire fonctionner votre outil.

Matériel :
Tout développement sera réalisé sur la plate-forme bio-informatique dédié à la génomique végétale dont l'infrastructure est la suivante :

- 2 serveurs de calcul de production (DellTM PowerEdgeTM 6650 - 4 processeurs Xeon 2.7 Ghz & 8 Go de RAM)
- 1 serveur de développement
- Un système de stockage RAID Dell/EMC (1 To de stockage)
- Un serveur de fichier

La configuration optimale du serveur (permettant e.g. son installation à Nouméa) sera définie avec le SIL après la finalisation de l'application

Outils et bases de données sur le serveur

- Blastall program (version 2.2.10)
- Gene ontology files (component.ontology, function.ontology, process.ontology, gene_ontology_edit.obo)
- Gene association files for gene ontology annotation (uniprot_sprot.dat, gene_association.Compugen_GenBank)
- go-show-paths-to-root.pl
- Matcher program
- Phred program (version 0.020425.c)
- StackPack program (version 2.2.0)
- Sequences databases for blast runs (nr, nt and swissprot)
- Vecscreen program
- Genbank
- Swiss-Prot
- TIGR Rice Database

Sur les postes des clients :

- Un navigateur Web (e.g. Mozilla/firefox ou IE)
- Un tableur pour charger les reports Excel (Excel ou Calc de Open Office)

54. Comment ces choix influeront sur l'appropriation de votre outil par les utilisateurs cibles ?

Le choix initial est d'utiliser des logiciels gratuits facilement accessibles et déjà bien testés pour leur robustesse afin de pouvoir diffuser l'outil EST-db le plus facilement et largement possible.

55. Justifier le choix de ces technologies (conformité à des référentiels, robustesse, pérennité, communauté de développeur importante…) :

Robustesse, gratuité des logiciels utilisés pour le pipeline, logiciels très utilisés.

Perl :

- Langage de référence en bioinformatique
- Langage spécialisé dans l'extraction de données (lecture de DB....)
- Perl possède un communauté importante de developpeurs
- Bibliothèque BioPerl et goPerl disponibles sur Internet

Données en entrée et en sortie

56. Énumérer et décrire les données en entrée et en sortie de votre outil.

Entrée : Chromatogramme de séquençage ou séquences en format FASTA

Sortie : liste de séquences annotées et référencées recherchées par mots clefs ou par blast (pages HTML); possibilité d'exporter des séquences sous format fasta

57. Décrire la disponibilité (ou l'accessibilité), le format de stockage et d'organisation ainsi que la qualité des données utilisées en entrée. Quel est le coût et l'effort requis de l'utilisateur pour collecter, acheter, obtenir ou convertir ces données ? Dans quelles mesures le coût et l'effort requis limiteront-ils l'adoption de votre outil ?

Les données de départ sont issues d'un séquençage et nécessite donc un budget spécifique pour cela. Côut indicatif : 30500 euros/ 10000 séquences en utilisant les ressources de la genopôle Languedoc-Roussillon qui pratique des tarifs plus bas que les sociétés privées. Certains des projets présentés ont financé leurs séquençage , d'autres ont pu obtenir des projet type Genoscope qui leur à permis d'obtenir leur séquence sans apporter un financement particulier.

Les prix sont de plus en plus accessibles et les projets sont de plus en plus nombreux. Notre outil sera vraisemblablement de plus en plus sollicité.

58. Les données seront-elles testées ou validées par l'outil en entrée ? Si oui, comment ?

Les fichiers bruts de chromatogrammes (ab1 ou scf) sont analysées en entrée du pipeline tout d'abord par le logiciel Phred ; si le format du fichier (extension, chimie etc.) n'est pas celui reconnu par Phred, ce fichier sera ignoré par le pipeline.

59. Validerez-vous ou avez-vous déjà validé scientifiquement les données en sortie de votre outil ? Si oui, décrire comment cela se fera ou a été fait.

Oui , Cf Liste des publications

60. Décrire l'utilité immédiate des données en sortie de votre outil et les nécessaires conversions, post-traitements ou analyses ultérieures requis. Comment l'effort requis impactera-t-il l'adoption de votre outil par les utilisateurs cibles ?

L'analyse des données issue du séquençage est un résultat publiable en soi. Des publications déjà réalisée le montrent. Dans un deuxième temps, la validation des résultats obtenu *in silico* est nécessaire par des techniques biologiques (Biologie moléculaire, ..).

Par ailleurs, la possibilité d'avoir en sortie les séquences sous format fasta est très utile car elle offre les possibilités de les traiter ultérieurement sans effort particulier (alignements contre des banques personnels, d'autres pipelines comme le SSR pipeline de recherche de microsatellites etc…)

61. Existent-ils des métadonnées ou y a-t-il production de métadonnées décrivant les lots de données en entrée ou sortie ? Si oui, comment sont-elles gérées et entreposées ? Sont-elles

basées sur des standards ?

Le pipeline utilise plusieurs bases de données standardisées disponibles sur Internet (ncbi, ebi, gene-ontology). Ces données seront mises à jour plusieurs fois dans l'années et sont directement stockées sur le serveur

62. La description ou le référencement des données est-il / sera-t-il basé sur un ou des référentiels ou thésaurus ? Si oui, lesquels ?

Le module "annotation Gene Ontology" est basé sur l'utilisation des standards d'ontologie biologique (OBO) :

L'annotation est basée sur la mise en rapport de différents fichiers (des fichiers faisant lien entre ID nt ou ID nr et ID GO, des fichiers .ontology et le fichier gene_ontology.obo) permettant à partir d'un pool d'EST d'annoter chaque séquence (lorsque cela est possible).

Interopérabilité

63. Quels sont les éventuels standards ou normes utilisées ?

Les standards W3C en termes de publication Web (XHTML, CSS).

L'ontologie biologique (Gene Ontology)

64. Votre outil est-il prévu pour être utilisé de manière interactive par les utilisateurs, par d'autres outils ou programmes (communication entre outils sur la base de requêtes ou autres) ou les deux ?

L'outil est utilisé d'une manière interactive par les utilisateurs (que ce soit les recherches dans la base de données ou encore le lancement du pipeline). Il ne sera pas utilisé par d'autres outils automatiquement (sauf, probablement, par le moteur de recherche du portail des ressources génomiques des caféiers ; cette intégration sera analysée dans un autre projet Spirales de la plateforme bioinformatique de Montpellier)

65. Si votre outil pourra être utilisé dans les 2 cas, de manière interactive et de manière automatisée par d'autres applications, décrire les caractéristiques et fonctionnalités non accessibles pour chaque mode d'utilisation.

Pas encore défini

66. Si votre outil pourra communiquer de manière automatisée avec d'autres programmes, écrire brièvement 3 scénarios d'utilisation qui illustrent les détails de ces communications.

Le pipeline d'analyse d'EST fait appel à plusieurs autres outils bioinformatiques cités ci-dessus (Phred, VecScreen, Matcher, Blast) d'une manière ordonnancée

67. Si votre outil intégrera ou fera appel à des outils d'autres développeurs, décrire brièvement 3 scénarios d'utilisation

Rapports d'erreurs et d'avancement

68. De quelle manière votre outil montrera la progression du traitement aux utilisateurs ? Qu'est-ce qui sera signalé ?

Un email est envoyé au départ du traitement à l'utilisateur, puis un autre une fois le traitement terminé. Entre temps rien ne permet à l'utilisateur de savoir où en est l'analyse.

69. Comment votre outil notifiera-t-il à l'utilisateur l'apparition d'une erreur et quelles informations seront affichées dans le message d'erreur ?

Les principales erreurs qui sont notifiées à l'utilisateur sont des erreurs pouvant apparaître au début de l'analyse (création des répertoires temporaires, et pour une analyse d'EST classique, récupération des chormatos). Si des problèmes sont rencontrés durant cette phase, un email sera envoyé à l'utilisateur pour lui dire la/les erreur(s) (exemple : ...Le repertoire X n'a pas pu être crée...Problème de copie des chromatos....). Pour faciliter l'intervention, les erreurs sont numérotées pour pouvoir se repérer chronologiquement dans l'enchaînement des opérations.

70. Avez-vous mis en place un processus de gestion des erreurs et de correction par l'équipe de développement et comment ?

Pour le moment, il n'y a pas de processus spécifique (hormis l'envoi de email) mis en place pour gérer les erreurs. Le seul moyen de savoir où le programme est bloqué c'est de regarder jusqu'ou les fichiers temporaires ont été générés. Par exemple si le blast de séquence à été effectué mais pas leur annotation on sait que le problème ce situe entre ces 2 étapes.

Documentation

71. Quelles sont les différentes documentations prévues : nature et format de la (des) documentation(s) ? cible visée ? (spécifications fonctionnelles, spécifications techniques, docs/API développeurs…)

La rédaction d'une documentation technique décrivant l'outil EST-db est un des objectifs de ce projet. Cette documentation servira de référence pour l'installation et l'utilisation de l'outil. Une notice aux utilisateurs est aussi en cours de rédaction.

72. Lister les sujets ou principaux chapitres qui apparaitront dans la/les documentation(s) de votre outil

Cf Document joint : Documentation technique.

Multilinguisme - traduction

73. Lister les langues parlées par vos utilisateurs cibles.

Français, Anglais, Espagnol, Portugais

74. Lister les langues dans lesquelles votre outil, votre documentation et tous les autres livrables seront traduits. Si vous ne traduisez par votre outil dans toutes les langues parlées par vos utilisateurs, comment cela affectera-t-il l'adoption de votre outil ?

Français - Anglais

75. Quelles méthodes ou technologies seront utilisées pour la traduction de votre outil, votre documentation et des autres livrables?

Rédaction directement en Anglais et validation après relecture par des anglophones

Processus et équipe de développement

76. Avez-vous déjà géré des projets de développement logiciel précédemment ? Décrire brièvement votre(vos) expérience(s) passée(s).

Oui. Deux autres projets de développement logiciel sont menés actuellement au sein de la plateforme

bioinformatique :

- Analyse, conception et développement d'un portail web dédié à la génomique du caféier au sein de la génopole de Montpellier, dans le cadre du réseau International Coffee Genome Network (ICGN) (support SPIRALES 2006 – 2007)
- InterProtDB, un système d'information dédié à la gestion et l'intégration de données protéomiques produites à haut débit (SPIRALES, début en 2007)

77. Les développements seront-ils réalisés par des membres de votre équipe, par un prestataire sous contrat, ou autre ?

Prestataire de service

78. Si vous avez déjà sélectionné des développeurs, de votre équipe ou d'un prestataire, lister les, spécifier leurs rôles et décrire leurs compétences et leurs expériences passées. Attacher leurs CV si vous les avez.

Advanced Solutions Accelerator, Castelnau le Lez porrait effectuer le travail. A confirmer si le projet est accepté .

79. Si vous envisagez un prestataire de service, avez-vous déjà travaillé avec un prestataire auparavant ? Décrire comment vous vous assurerez qu'il développe ce que vous recherchez, dans les temps et avec le budget prévu.

Oui, un projet Spirale antérieur a été réalisé avec ASA (Cf point 76)

Assurance pour le développement du projet :

Cahier des charge détaillé

Réunions « bilan » bimensuelles avec les utilisateurs et la responsable du projet bioinformatique (C. Dubreuil-Tranchant ou remplaçant)

Suivi régulier de l'animateur scientifique (V. Hocher)

Rédaction de compte rendu

Test par les utilisateurs

80. Impliquerez-vous vos utilisateurs cibles dans le processus de conception et d'implémentation de l'outil ? Si oui, décrire comment.

Un groupe de travail impliquant les différents scientifiques demandeurs a été créé et se réuni au moins une fois / mois avec la société prestataire de service. Des réunion ponctuelles sont par ailleurs organisé en cas de nécessité pour décidé d'une orientation a prendre.

A l'issue de la conception des nouveaux modules les utilisateurs auront une démonstration ainsi qu'une période de test afin de valider la fonctionnalité des nouveaux modules avant leur transfert sur la chaîne de traitement.

81. Où sera hébergé le code source de votre outil durant son développement puis durant sa maintenance ?

Hébergement durant le développement : serveur de test http://bio-info/estdb

Hébergement durant le fonctionnement: serveur de production http://bioinfo/estdb

82. L'outil sera-t-il placé dans une plateforme collaborative ou au sein d'une communauté de développement de projets open-source ? si oui, lesquels ?

Non défini pour le moment.

Licence et distribution

83. L'utilisation de l'outil sera-t-elle soumise à une licence pour les utilisateurs qui l'installeront sur leurs propres machines? S'agira-t-il d'une licence libre ? Le code source de l'outil sera-t-il protégé ou complètement ouvert ? (décrire l'éventuel coût, le type de licence et toutes autres éventuelles obligations)

Non, seule la signature d'une charte sera demandée. Cette charte stipulera l'origine ainsi que la liste des concepteurs de l'outil et la propriété intellectuelle. Ce point sera cependant à préciser avec C. Dubreuil-Tranchant.

84. Existe-t-il des parties ou modules de votre outil qui sont protégés par des brevets ou des marques ?

Non

85. Décrire comment l'outil sera distribué ou rendu accessible aux utilisateurs (lister les sites web si nécessaire)

L'application EST-dbsera intégrée dans le site Web de la plateforme bioinformatique de l'IRD (http://www.mpl.ird/bioinfo). Les données de la base de données EST-dbseront consultables sur le web . A la demande que se fera par email, les codes sources seront mis à la disposition des demandeurs sous forme de fichier compressé, avec une notice d'installation

Installation

86. La procédure d'installation sera-t-elle automatisée par un programme ou un script ou l'outil devra-t-il être installé « manuellement » ? (Préciser les OS et distribution)

L'outil devra être installé de manière manuelle (en suivant la documentation créee à cet effet).
OS : linux
Distribution : non précisée

87. Est-ce que le programme ou script d'installation détectera et signalera les logiciels requis manquants ?

Non car pas de programme d'installation automatique

88. Est-ce que le programme ou script d'installation permettra la désinstallation de l'outil ?

Non car pas de programme d'installation automatique

89. Si l'installation n'est pas pris en charge par un programme ou un script, existera-t-il une notice d'installation ?

Oui c'est en cours de rédaction, ce point fait partie intégrale du projet .

90. De quelle manière la complexité de la procédure d'installation limitera l'adoption/l'utilisation de l'outil par les utilisateurs cibles ?

Bien que manuelle, la procédure d'installation n'est pas complexe ; accompagnée d'une notice d'installation, elle ne posera pas de soucis pour l'adoption de l'outil par les utilisateurs cibles. Toutefois, l'installation devra être effectuée par un informaticien / bioinformaticien possédant les droits root sur le serveur.

🗀 Opération

91. Les utilisateurs pourront-ils faire fonctionner l'outil sans votre aide? Si les utilisateurs doivent solliciter votre équipe ou des consultants externes ou suivre une formation, décrire les détails et les coûts

Oui, il y aura une possibilité d'aide. Une formation de courte durée sur l'utilisation du pipeline, des fonctionnalités proposées par l'application peut être envisagée. Ce sera réfléchi en fonction de la demande des utilisateurs et surtout géré par C. Dubreuil-Tranchant..

🗀 Assurance qualité, maintenance et support

92. Lister les techniques que votre équipe utilisera pour détecter les erreurs ou défauts.

Pour connaître les éventuels problèmes de l'application, on va stocker les messages d'erreur dans un fichier de log d'EST-db. L'administrateur sera également averti par email en cas de problèmes graves. Le fichier log d'apache sera également examiné en cas de problème.

93. Dans le cas où vous auriez un programme 'beta' en fin de développement, décrire comment il fonctionnera. Si des utilisateurs se sont déjà engagés pour l'utiliser, listez-les.

94. De quelle manière votre équipe fera-t-elle le suivi des erreurs dans ce projet ?

Toutes les erreurs du projet, comme toutes les taches accomplies avec succès, seront notées dans le fichier du suivi de projet. Les erreurs seront signalées au chef de projet. En cas d'urgence, une réunion de crise sera organisée afin de corriger le problème créé

95. De quelle manière apporterez-vous un support à vos utilisateurs pendant la durée de ce projet, et après ?

Pendant la durée du projet, les utilisateurs continuent à utiliser l'ancienne version de EST-db qui possède les fonctionnalités basiques pour leurs analyses.

Après la mise en place de la version améliorée, le suivi / entretien / mise à jour : gestion des erreurs…. Sera assuré par le responsable de la plateforme bioinformatique C Dubreuil-tranchant ou remplaçant.

En tant qu'administrateur de bases de données, il assurera la maintenance de la base de données (amélioration des performances, de sécurité, gestion d'erreurs, sauvegardes, mises à jour...)

96. De quelle manière apporterez-vous un appui aux développeurs d'autres outils qui souhaiteraient utiliser et intégrer votre outil aux leurs ?

En cas de demande d'intégration de notre outil dans un outil externe, après un accord préalable entre les chefs des projets et les responsables scientifiques d'EST-db, nous allons mettre à la disposition de ces développeurs la documentation technique et le schéma de la base de données et du pipeline.

7. Description du projet – HORS développement d'application IS (méthodes, solutions, et moyens)

Cette partie (questions 97 à 99) concerne les **nouveaux projets ainsi que les demandes de continuums (HORS développement d'application IS).**

La demande peut être **accompagnée de tous documents utiles** :

présentation du projet global ou descriptif du projet, rapport de phases

préliminaires, étude de faisabilité, dossier d'expression des besoins ou ***cahier des charges****, devis détaillé…*

97. Description du projet :

98. Description technique du projet / choix technologiques (*si approprié*).

99. Organisation, faisabilité et échéancier du projet.

8. Pertinence, résultats/livrables attendus et valorisation du projet

Cette partie (questions 100 à 106) doit être renseignée **quelque soit la nature de la proposition** (*nouveau projet ou continuum d'un projet SPIRALES existant, étude de faisabilité, projet finalisé de développement d'une application IS ou autre*).

La demande peut-être être **accompagnée de tous documents utiles** :

présentation du projet global ou descriptif du projet, rapport de phases préliminaires, étude de faisabilité, dossier d'expression des besoins ou ***cahier des charges****, devis détaillé…*

100. Résultats attendus (livrables) : (10 lignes maximum)

- Version 1.0 d'EST-db documentée, suffisamment générique pour être installée sur n'importe quel serveur bio-informatique
- Distribution d'EST-db aux équipes IRD ou partenaires intéressés pour utiliser l'application après déclaration à l'APL.
- Développement de nouvelles fonctionnalités ajoutant une valeur ajoutée importante à EST-db.
    - Possibilité de lancer le pipeline par les partenaires et IRDiens expatriés
    - Lancement du pipeline à partir des séquences (format fasta) et pas uniquement des chromatogrammes
    - Amélioration des interfaces de consultation
    - Mise en place d'un module de visualisation des séquences ESTs comparées à leur contig
    - Amélioration de l'interface d'administration.
- Module d'Annotation Automatique en grande fonction.
- Module analyse comparative inter et intra espèces
- Rédaction d'une publication dans la revue appropriée et/ou d'une communication dans un congrès international

101. Pertinence du projet pour votre communauté scientifique

Les équipes IRD concernées par ce projet utilisent déjà en routine l'application EST-db pour traiter leurs données de génomique. Ce projet, s'il continue, permettra de réaliser de nouveaux types d'analyses indispensables comprendre la biodiversité et rechercher de nouveaux gènes d'intérêts chez les plantes tropicales. EST-db est un outil incontournable pour la réalisation de nos programmes de recherche et pour leur valorisation sous forme de publication. Ce projet SPIRALE permettra aussi l'optimisation d'EST-db afin de rendre cette application transférable et utilisable pour les équipes IRD expatriées et les partenaires extérieurs.

102. Pertinence du projet vis à vis des objectifs de SPIRALES / justification d'un financement DSI

Le projet que nous proposons entre tout à fait dans les objectifs de SPIRALES. Il s'agit :

de finaliser l'application EST-db utilisée par 4 UMRs différentes en ajoutant de nouvelles fonctionnalités et en produisant un code suffisamment paramétrable afin que l'application soit installée sur n'importe quelle plate-forme et distribuée à des partenaires.

de développer de nouveaux modules à EST-db, outil conçu, développé et utilisé de façon transversale par les différentes équipes/UMRs plantes présentes à l'IRD. Ce projet qui concerne plus spécifiquement la génomique comparative permettra d'étendre l'utilisation d'EST-db à des partenaires extérieurs d'organismes tels que l'INRA (Cf Programme Biodiversité) ou le CNRS (Cf Programmes Symbioses actinorhiziennes) ainsi qu'à des partenaires du Sud (Cf Programme Hévéa).

103. Retours sur investissement attendus (pour l'unité, l'institut…)

- Pour les unités :
    - Analyse et exploitation des données génomiques
    - Avancée dans les connaissances fondamentales en biologie des plantes
    - Réalisation de publications.
- Pour l'institut :
    - Développement d'une application (EST-db) générique transférable
    - Développement de la première plateforme bio-informatique IRD
    - Développement de compétences nouvelles en bio-informatique
    - Expertise IRD en bio-informatique et transfert aux partenaires

104. Capitalisation, valorisation, transfert de savoirs-faire ou d'outils possibles ou prévus en matière d'IS

Toute la méthodologie sera décrite et l'application ainsi que la documentation associée (rapport de sage, documentation technique) sera transmise sur demande au sein de l'institut en vue de transférer notre savoir faire et expertise dans ce domaine.

La valorisation de cet outil sera aussi assurée par le biais de publications dans des revues spécialisées, de communication à des congrès, de formations organisées pour les équipes potentiellement intéressée.

Les chercheurs associés à ce projet valoriseront l'application EST-db en publiant les résultats issus des analyses réalisées dans les revues scientifiques adéquates.

105. Valorisation possible ou prévue

- Rédaction d'une publication dans la revue appropriée et/ou d'une communication dans un congrès international
- Déclaration de l'application EST-db à l'APL
- Distribution d'EST-db aux équipes IRD ou partenaires intéressés et mise en ligne au niveau du site de la plateforme bio-informatique (http://www.mpl.ird.fr/bioinfo) de la nouvelle version d'EST-db
- Diffusion de la méthodologie (cahier des charges) aux partenaires et au sein de l'IRD du code des nouveaux modules « Annotation en grande fonction » et « Génomique comparative», analyses réalisées par toutes UR développant des projets de génomique (domaine animal ou végétal).
- Analyse des données biologiques plantes disponibles et publications des résultats des analyses dans les revues adéquates par les chercheurs.

106. Observations particulières :

Remarque 1 :

La somme demandée pour 2008 (20000 € HT) est supérieure à celle que nous avions estimée en 2007. En effet, compte tenu du cahier des charges initial, du temps nécessaire pour développer les outils (temps que nous avions sous-estimé en 2007), nous avons ré-estimé le coût et décidé de faire une demande correspondant à l'obtention du produit fini transférable. Par ailleurs, il est important de remarquer que les UMRs concernées par ce projet ont apporté 10000 euros en 2007 et apporteront une somme identique si les budgets 2008 sont validés. Enfin, un recrutement bioinformatique a été demandé pour renforcer le personnel

de la plateforme. L'ensemble de cette dynamique montre l'importance de cette plateforme pour les UMRs plantes et la nécessité de disposer de cet outil et de pouvoir le transférer auprès de nos partenaires et plus largement si nécessaire.

Remarque 2

C. Dubreuil-Tranchant étant en congé maternité, nous avons rédigé ce compte rendu et ce projet sans elle ; Certains détails n'ont donc pu être renseignés, mais dès son retour elle pourra fournir toutes les informations supplémentaires si nécessaire.